ACCURACY OF ITEM RESPONSE THEORY PARAMETER ESTIMATES USING MAXIMUM LIKELIHOOD AND BAYESIAN PROCEDURES AS IMPLEMENTED IN LOGIST AND BILOG

BY

ABDEL-FATTAH A. ABDEL-FATTAH

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN
PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1990

ACKNOWLEDGMENTS

Thanks are expressed to my chairman, James Algina, who contributed to the progress of my dissertation. I also extend my gratitude to the members of my doctoral committee, Linda M. Crocker, Fatima Linda C. Jackson, and David Miller for their cooperation.

TABLE OF CONTENTS

Abstract of Dissertation Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirement for the Degree of Doctor of Philosophy

ACCURACY OF ITEM RESPONSE THEORY PARAMETER ESTIMATES USING MAXIMUM LIKELIHOOD AND BAYESIAN PROCEDURES AS IMPLEMENTED IN LOGIST AND BILOG

By

Abdel-fattah A. Abdel-fattah

August, 1990

Chairman: James Algina
Major Department: Foundations of Education

The purpose of this study was to compare the accuracy of three estimation procedures in item response theory: the joint maximum likelihood as implemented in the computer program LOGIST, the marginal maximum likelihood, and the marginal Bayesian procedures as implemented in the computer program BILOG.

The comparisons were conducted using data generated by a Monte Carlo simulation based on the three-parameter logistic model. The data characteristics varied in each simulation were the number of items, the number of subjects, and the distribution of ability parameters. The ability parameter distribution was the variable of most concern.

Normal ability distributions provided more accurate parameter estimates with respect to the Marginal Bayesian estimation procedure, especially when number of items and

number of examinees were small. The Marginal Bayesian estimation procedure was generally more accurate than the other two procedures in estimating a, b, and c parameters. When the ability distribution was beta the Joint Maximum Likelihood was the most accurate in estimating the c parameters.

Guidelines were provided for obtaining accurate estimation using real data and sample sizes, test lengths, and ability parameter distributions investigated in this dissertation. For example, the Marginal Bayesian procedure is recommended with short tests and small samples for estimating a, b, and c parameters. The Joint Maximum Likelihood is preferred when guessing is a problem of main concern and the ability distribution is beta.

CHAPTER I
INTRODUCTION

Overview

The purpose of this study was to investigate the accuracy of three of the most important parameter estimation procedures in item response theory (IRT): joint maximum likelihood, marginal maximum likelihood, and marginal Bayesian. These procedures were investigated under variations of several factors that affect estimation accuracy: ability parameter distribution, sample size, and test length. This chapter includes

1. an introduction to binary IRT and its basic assumptions,
2. an introduction to the estimation of the IRT parameters,
3. a statement of the purpose of the study,
4. a brief description of methodology, and
5. a description of the significance of the study.

Binary IRT Models

Binary item response theory is a theoretical framework for modeling scores on dichotomous items, such as those found on typical ability and achievement tests. Several IRT models have been proposed for an item characteristic curve (ICC), the relationship between the ability (θ) of examinees,

and the probability that item i is answered correctly. One of these models is the three-parameter logistic model. The mathematical form of the three-parameter logistic curve is

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \tag{1}$$

where $P_i(\theta)$ is the probability that an examinee with level θ answers item i correctly, a_i is the discrimination level of item i, b_i is the difficulty of item i, c_i is the pseudo chance-level parameter of item i, and D is a scaling factor.

Birnbaum (1957, 1968) proposed an item response model in which the ICCs take the form of two-parameter cumulative logistic distribution functions:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \tag{2}$$

An inspection of the two-parameter logistic model reveals an implicit assumption that is characteristic of many item response models: $c_i = 0$.

In the one-parameter logistic model, another special case of Birnbaum's three-parameter logistic model, $c_i = 0$ and a_i is a constant for all items. The ICC for the one-

parameter logistic model can be written as

$$P_i(\theta) = \frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}} \qquad (3)$$

in which "a" is the constant level of discrimination for all items.

The basic assumption of IRT is local independence: Each examinee in a population may be characterized by his or her score on one or more latent variables such that in a subpopulation of examinees, each with the same score on each latent variable, responses to the items in the test are mutually statistically independent. With the preceding IRT models, it is assumed that the items are unidimensional. That is, a single latent variable exists such that the item responses are locally independent. With these models the local independence assumption can be expressed mathematically as follows:

$$P(\underline{U}|\theta) = \prod_{i=1}^{n} [P_i\ (\theta)]^{u_i}\ [1-P_i\ (\theta)]^{1 - u_i} \qquad (4)$$

where $\underline{U}$ is the random item response pattern vector, u_i is the dichotomous random response variable on the ith item, θ is the random latent variable, $P(\underline{U}|\theta)$ is the probability of

the response pattern conditional upon the latent random variable, and $P_i(\theta)$ is the probability of responding correctly to item i conditional upon θ. The assumptions required by the IRT models are thus local independence and the functional form for the ICC.

Estimation of Parameters

Once the data (the scores on the n items) are observed, equation (4) ceases to be interpretable as a probability and becomes the likelihood function for examinee j. The likelihood function for a sample of N examinees is the product of the individual likelihoods. A standard estimation strategy is to maximize the likelihood function or some variant of the likelihood function. Several such procedures are available. These procedures, which will be presented in the second chapter, are the joint maximum likelihood (JML) procedure, the marginal maximum likelihood (MML) procedure, and the Bayesian versions of these procedures. We shall refer to these as the joint Bayesian (JB) and the marginal Bayesian (MB) procedures. The JML and the MML are implemented in the computer programs LOGIST (Wingersky & Lord, 1973; Wingersky, Barton, & Lord, 1982; Wingersky, 1983) and BILOG (Mislevy & Bock, 1984) respectively. The MB procedure is implemented in BILOG and the JB is implemented in a program constructed by

Swaminathan and Gifford (1986). Provision for general distribution of the latter program has not been made. The JML procedure makes no assumptions about item or ability parameter distributions, the MML procedure requires an assumption concerning ability distribution, and the JB and the MB procedures make assumptions about the distribution of both item and ability parameters.

Purpose of the Study

The purpose of this research was to compare accuracy of estimation using the three estimation procedures: JML, MML, and MB. Estimation accuracy of the parameters in the three-parameter model was investigated. This accuracy was investigated both when the normal ability assumption of the MML and MB was met and when it was violated. Three ability distributions were used: normal, truncated normal, and beta. The specific conditions investigated are described in Chapter III.

Methodology

There are two possible empirical approaches to investigate accuracy of parameter estimation in IRT. One approach is to use real data and the other is to use simulated data. The usefulness of real data is limited because the true parameter values are unknown. The prime

advantage of using the simulated data is that the parameters are known. In addition, parameter values can be manipulated so that a reasonably broad set of conditions can be investigated. The major drawback of using simulated data is that the conditions investigated may not correspond exactly to conditions found in real data. In this study the accuracy of the estimation procedures was investigated by using simulated data.

Significance of the Study

Item response theory is used in scoring tests, equating tests, investigating item bias, cross cultural research on tests, establishing item banks, and investigating validity. Consequently, comparing the accuracy with which available procedures, such as BILOG and LOGIST, estimate item and ability parameters is critical to the use of IRT.

Mislevy and Stocking (1987) indicated that one path to compare fairly BILOG and LOGIST was to broaden the range of generated values of parameters and that this path has been led by Yen (1984, 1987) who compared the two programs over a range of generated values broader than previous research. The three ability distributions used in this dissertation is an example of extending the range of generated values beyond previous research (e.g., Ree, 1979).

CHAPTER II
REVIEW

Three objectives are addressed in this chapter. The first is to present the four estimation methods identified in the previous chapter. The second is to review the literature that compares the accuracy of the four types of estimators as sample size and test length vary. The third is to review the literature related to distributional variations in IRT parameter estimation.

Estimation Methods

The basic problem in item response theory is to estimate the parameters a_i, b_i, c_i, and θ_j that characterize the model. There are currently four major procedures that are statistically well founded and can be used with all three logistic models: the JML, the MML, the JB, and the MB methods.

The JML Procedure

As noted in the previous chapter, the likelihood function for the N examinees is the product of N likelihood functions given by equation (5):

$$L(\theta;a,b,c) = \prod_{j=1}^{N} \prod_{i=1}^{n} [P_i(\theta_j)]^{u_{ij}} [Q_i(\theta_j)]^{1-u_{ij}} \qquad (5)$$

where $Q_i(\theta_j) = 1 - P_i(\theta_j)$ and $P_i(\theta_j)$ is defined by equations (1), (2) or (3) as appropriate. The JML procedure consists of finding estimates of θ_j, a_i, b_i, and c_i that maximize the likelihood function or equivalently the estimates that maximize the logarithm of the likelihood function:

$$L = \sum_{j=1}^{N} \sum_{i=1}^{n} [u_{ij} \log P_i(\theta_j) + (1-u_{ij}) \log Q_i(\theta_j)] \qquad (6)$$

where L is equal to log $L(\theta,a,b,c)$.

An iterative procedure is carried out in two stages. The first stage starts with the initial values of a_i, b_i, and c_i, treating θ_j as unknown. Once the θ_js are estimated, the second stage starts with the θ estimates from the first stage and treats the item parameters as unknowns to be estimated. This two-stage process is repeated until the ability and item values converge, with the final values being taken as the JML estimates.

The major problem with the JML is that when the items are viewed as fixed, rather than as a random sample from a universe of items, estimates of the item parameters are not consistent. That is, the estimates do not converge to their true values even as the number of examinees goes to infinity. This is because as N goes to infinity there are infinitely many θ_js to estimate; this has negative consequences for estimating the item parameters. However,

Haberman (1975) has shown that as both n and N go to infinity, the JML estimators in the one-parameter model are consistent. Empirical results reported by Lord (1975) and Swaminathan and Gifford (1983) suggest this is true for the two- and the three-parameter models also.

The program LOGIST is based on the JML procedure developed by Lord (1974). It has been available since 1973 (Wingersky & Lord, 1973) and has undergone major revision (Wingersky, 1983; Wingersky, Barton & Lord, 1982). Initial values for the item parameter estimates are set, and ability estimates are computed for all the examinees using the maximum likelihood estimation. Then the ability estimates are held fixed, and new estimates are made for the a_i, the b_i, and the c_i values, again using the maximum likelihood estimation. The procedure cycles through as many stages as necessary to convergence. Convergence is achieved when the difference between the estimates of successive stages is negligible.

To keep the estimates of a_i and c_i from drifting out of bounds, constraints are placed on the estimates of a_i and c_i. An upper limit is placed on the estimated discrimination parameter. The default upper limit is 2.0. In the initial iterations, the c_i values are held fixed. After the first few stages, the c_i values are allowed to vary, but changes in the c_i values are still restricted.

Swaminathan and Gifford (1987) indicated that this procedure has been criticized, because the estimates obtained in this way may not be true maximum likelihood estimates. In spite of this criticism, Swaminathan and Gifford placed an upper limit of 2.0 and a lower limit of 0.06 on the a parameter because they were more interested, as it seems, in comparing LOGIST (not the pure JML) to BILOG.

The JB Procedure

In the JB methods (Swaminathan & Gifford, 1982, 1985, 1986) the likelihood in equation (5) is multiplied by a prior for each of the item and the ability parameters to obtain an expression proportional to the joint posterior distribution of these parameters. The JB function is

$$f(\theta;a,b,c) = L(\theta;a,b,c)g(\theta)g(a)g(b)g(c) \tag{7}$$

Swaminathan and Gifford used a normal/normal/gamma/beta prior for the θ_j, b_i, a_i, and c_i parameters. They used these priors, assuming independence of their parameters, to compute joint modal estimates of item and ability parameters. The use of the prior distributions tends to prevent parameter estimates from drifting to intuitively unreasonable values. The JB procedure is hierarchical with respect to b and θ parameters but non-hierarchical with respect to a and c parameters. Swaminathan and Gifford

(1982, 1985, 1986) have implemented the JB procedure in a computer program that is not currently available for general distribution.

ASCAL is a microcomputer-based program that implements a modified JB procedure for the three-parameter logistic model. The procedure used was modeled after the JML procedure of LOGIST combined with a modal Bayesian procedure (Vale & Gialluca, 1985). In ASCAL the likelihood equations, which were modified for omitted items (Lord, 1974), were combined with Bayesian prior distributions on the a, c, and θ parameters.

These equations were modified to take into account a normal Bayesian prior distribution on the ability parameters and beta prior distributions on the a and c parameters. For the a parameters the Bayesian prior is a beta distribution with parameters $R = S = 3.0$, $L = 0.3$, and $U = 2.6$, where R and S are the shape parameters and L and U are the lower and the upper limits respectively. Estimates of the a parameters are bounded at 0.4 and 2.5. For the c parameter the Bayesian prior is a beta distribution with $R = S = 5$, $L = -0.05$, and $U = 2/K + 0.05$, where K is the number of alternatives. The c estimates are bounded at 0 and $2/K$. The c parameters are estimated using the same modal Bayesian procedure used for the a parameters. The b parameters are computed by using the JML procedure with bounds -3 and 3.

A modified multivariate Newton-Raphson procedure is used in the item parameter estimation of ASCAL. The estimation process begins by specifying starting points for the ability and item parameters. These points are the estimates implemented in early versions of the computer program ANCILLES and they are calculated by an heuristic approximation procedure (Jensema, 1976). The abilities are then estimated using the initial values of the item parameters. The estimated abilities are sorted into 20 fractiles with approximately equal number of examinees in each. The item parameters are then estimated by using the 20 fractile means instead of the entire ability distribution. The sequence of ability estimation, ability grouping, and item parameter estimation is repeated until ability and item parameter estimates converge on stable values or fail to improve. ASCAL is designed to estimate only item parameters and is not intended to provide ability estimates.

The MML Procedure

The MML procedure was introduced by Bock and Lieberman (1970). The problem of inconsistent item parameter estimators is eliminated in the MML procedures by obtaining the marginal, rather than the likelihood function that is conditional on the ability parameters. Multiplying equation (4) by $g(\theta)$, the probability density function for the

ability parameters, and integrating with respect to θ we obtain the marginal probability of the response pattern $\underline{U}$

$$P(\underline{U}) = \int_{-\infty}^{\infty} P(\underline{U})|\theta)g(\theta)\ d\theta \tag{8}$$

Again, once the data are observed this probability can be interpreted as marginal likelihood function for a particular examinee. The product of these likelihoods for all examinees yields the marginal likelihood function for the entire data set. The marginal likelihood function for all examinees is

$$L(a,b,c) = \prod_{j=1}^{N} \int_{\theta} L(\theta;a,b,c)g(\theta)\ d\theta \tag{9}$$

The MML estimates are the estimates of a_i, b_i, and c_i that maximize this likelihood function.

Bock and Lieberman (1970) gave a numerical solution to the likelihood equations. The solution was computationally burdensome and its application was limited to tests with 10 or fewer items. Bock and Aitkin (1981) reformulated the likelihood equations of the Bock and Lieberman solution to produce a solution that avoids these computational problems.

The MML procedure avoids the problem of estimating θ for each subject. Therefore, it is intuitively clear that the MML is especially advantageous for short tests. MML item parameter estimates are consistent for tests of any length (Bock & Aitkin, 1981) as the number of subjects increases. However, the a_i parameter estimates in the two- and the a_i and c_i parameter estimates in the three-parameter logistic models may drift to extreme values. BILOG (Mislevy & Bock, 1984) is a recently developed program that produces MML estimates.

<u>The MB Procedure</u>

In the MB procedure the likelihood given by equation (9) is multiplied by prior distributions for a, b, and c. The resulting expression, $L(a,b,c)g(a)g(b)g(c)$, is proportional to the posterior density for a, b, and c.

One advantage of the use of the MB estimation is its tendency to prevent item parameter estimates from drifting to extreme values. The extreme values are pulled toward the center of the prior distribution for the item parameters, whereas that center differs a little from where it would have been without the use of the priors (Mislevy & Bock, 1984).

BILOG has the capability of MB estimation. The MB procedure in BILOG can be used with fixed or floating priors. With fixed priors, the means of the prior

distribution remain the same at each iteration. With floating priors, the means vary over each iteration. At each iteration, the mean of the prior for a particular parameter (a, b, or c) is set equal to the mean estimated value of that parameter from the previous iteration. According to Mislevy and Bock (1984) this is tantamount to hierarchical Bayesian estimation. The default priors used in BILOG are normal for the θ parameters, lognormal for the a parameters, normal for the b parameters, and beta for the c parameters.

Related work on the MB procedure can be found in Dempster, Rubin, and Tsutakawa (1981), Rigdon and Tsutakawa (1983, 1987), and Tsutakawa (1984, 1986). The iterative solution, introduced by Dempster et al. (1981), was more general than the similar solution by Bock and Aitkin (1981). The Bock and Aitkin solution was limited to random variables with exponential distributions. This limited solution was extended by Rigdon and Tsutakawa to employ random variables belonging to non-exponential family distributions.

Using the Rasch model, Rigdon and Tsutakawa (1983) derived a marginal maximum likelihood with a fixed difficulty parameter and a random ability parameter. This procedure is called the maximum likelihood fixed (MLF) procedure. The ability parameters were assumed to be sampled from a population distribution, which was selected

from a prior distribution. The true likelihood function is the integral of the conventional likelihood function with respect to ability distribution. The resulting likelihood function is the function of item parameters and the prior distribution of the unknown ability distribution. This extended solution is thus a refinement of earlier applications (e.g., Bock & Aitkin, 1981).

The conditional maximum likelihood fixed (CMLF) procedure was developed from the MLF by using the posterior mean of each examinee's θ in the estimation process of the priors. The CMLF procedure approximates the conventional estimation of the unknown Bayesian priors (e.g., θ or b priors) conditioned upon the posterior mean of this prior. This approximation reduces the number of computations required by the conventional MML procedure when it is used in estimating the prior distribution.

Rigdon and Tsutakawa (1987) derived two more MB procedures under the Rasch model. These procedures are called the conditional maximum likelihood random (CMLR) and the conditional maximum likelihood uniform (CMLU). The prior distribution of the b parameter was random in the CMLR procedure and uniform in the CMLU procedure. The ability parameters were assumed random with a normal prior distribution. The two procedures are thus fully Bayesian extensions of the CMLF procedure because each of θ and b

parameters has certain prior distribution. Rigdon and Tsutakawa have implemented their MB procedure in a computer program. This program is not only unavailable for general distribution but also is restricted to the one- and the two-parameter logistic models.

Sample Size, Test Length, and Estimation Procedure

Urry's Method and the JML Procedure

Swaminathan and Gifford (1983) investigated the accuracy of parameter estimation in the three-parameter model using the modified JML procedure implemented by LOGIST and an approximate estimation procedure implemented in the ANCILLES program (Urry, 1977). The factors and levels in the design are reported in Table 1; the design was completely crossed.

In general, the results indicate that the JML procedure was superior to the Urry method with respect to estimation of all item and ability parameters, especially in the case of short tests. The difference between the two procedures became negligible as the number of items and the number of examinees increased. However, ANCILLES required considerably less computer time than LOGIST. The rapidity of convergence in ANCILLES was due to the fact that it deletes more items and examinees during the estimation procedure than LOGIST. For the JML procedure, differences

TABLE 1

Factors and Levels in Swaminathan and Gifford (1983)

Factors	Levels
Number of Items	10, 15, 20, 80
Number of Examinees	50, 200, 1000
Parameter Distributions[a]	
θ	Normal, uniform, negatively skewed[b]
	(-1.73, 1.73)[c]
a	Uniform
	(0.6, 2.0)
b	Uniform
	(-2, 2)
c	Uniform
	All c=.25
Estimation Procedure	JML (LOGIST) and Approximate (ANCILLES)

[a]The three-parameter model was studied.
[b]The skewed ability was generated from a beta distribution with scale parameters 5 and 1.5.
[c]The figures in parentheses give the range of potential values for the parameters.

between true and estimated parameters decreased with increases in both sample size and number of items. This trend was more obvious for the a and c than for the θ or b parameters. The number of examinees had a slight effect on improving the accuracy of estimation of the b, c, and θ parameters. Increasing the number of items and the number of examinees, however, considerably improved the accuracy of the a estimates with both procedures. A 20-item test with 1000 examinees produced excellent estimates of the b and c parameters and reasonably good estimates of the a and θ parameters. Tests with 80 items and 1000 examinees provided good estimates of all parameters.

Hulin, Lissak, and Drasgow (1982) investigated the accuracy of the JML parameter estimation in the two- and the three-parameter logistic models. Table 2 shows the factors and levels used in the study.

The correlations for the difficulty parameters with their estimates, in the two- or the three-parameter logistic models, were substantially lower when a difficulty range of (-3, +3) was used. When the difficulty range was (-2, +2), the same range used by Swaminathan and Gifford, the results were consistent with the results reported by Swaminathan and Gifford (1982). To avoid the adverse effect produced by the higher range, we used a range similar to the small range as suggested by the review above.

TABLE 2

Factors and Levels in Hulin, Lissak, and Drasgow (1982)

Factors	Levels
Number of Items	15, 30, 60
Number of Examinees	200, 500, 1000, 2000
Parameter Distributions	
θ	Standard Normal
a	Positively Skewed
	(0.19, 1.60)[a,b]
b	Uniform
	(-3, 3) and (-2, 2)
c	Uniform
	(.11, .33)
Models	Two- and Three-parameter

Note. JML implemented by LOGIST was studied.
[a]The figures in parentheses give the range of potential values for the parameters.
[b]Mean = 0.862, standard deviation = 0.209.

The accuracy with which the a and b parameters were estimated by JML was very similar for the two-parameter and the three-parameter models. An important result was that above the level of 1000 subjects and 60 items, there was no noticeable increase in accuracy of the a or the b parameter estimation for either the two- or the three-parameter models. The above results were thus in general agreement with Lord's (1968) conjecture that as many as 50 items and 1000 examinees may be required for accurate estimation of the a parameter in the three-parameter logistic model.

Wingersky and Lord (1984) studied the sampling errors of the JML estimates under the three-parameter logistic model. The responses of 1500 examinees on a 50-item science test were calibrated by the JML of LOGIST to obtain the item and ability parameter estimates. From the 50-item test, a sample of 15 items with their parameter estimates was drawn. The 15 items with their parameter estimates were repeated three times to form a 45-item test. A 90-item test was created by duplicating the 45-item test. In each of the 45-item and the 90-item tests the 1500 θ parameters were then replicated four times to represent the 6000 θ parameters. The distributions of the θ parameters were normal in both of the 45- and the 90-item tests. For the 45-item test a random sample of 1500 θ parameters was drawn from a rectangular distribution in the range (-3, 3).

The standard errors of the first 15 items were calculated for item parameters of the 45- and the 90-item tests. The abilities were grouped into 16 intervals; then the standard error was calculated for each interval.

Wingersky and Lord reported that quadrupling the number of examinees reduced the standard errors of item parameters by half and sharply reduced the standard errors of the largest ability estimates. Doubling the number of items decreased the standard errors of the abilities by a factor of 2.5 and either had a moderate or a little effect on the standard errors of item parameters. Therefore they concluded that increasing the number of items and the number of examinees reduces the standard errors of item parameters. This reduction in the standard errors is an indication of convergence of JML estimates of the three-parameter model to the true values with the increase of sample size and test length.

Repeating the item and/or ability parameters, to obtain large sample sizes and test lengths, seems to control for factors affecting accuracy of estimation except sample size, test length, and the distribution of the unrepeated parameters. A similar approach was used in this dissertation. The θ parameters of the small sample size are contained in those of the large sample size, and the item parameters of the short tests were subsets of the long

tests. The use of LOGIST estimates as the true values may underestimate the standard errors so that the JML estimation, of its own estimates, appear to be very accurate. The true values in this study were not estimates of either LOGIST or BILOG.

The JML and the MML procedures

Swaminathan and Gifford (1987) have recently compared MML and JML for the one-, two- and three-parameter logistic models. The design of the study is shown in Table 3. The test length, sample size, program, and model factors were completely crossed.

The JML estimator was again found to be ineffective with short tests and small samples of examinees. Item parameter estimates did not appear to converge to the true values, for a fixed test length as the sample size increases. MML, as mentioned earlier, possesses this property of conversion to the true value as sample size increases. Both procedures were again found to result in estimates that converge to the true values when both the number of items and the number of examinees increase.

The results show that with the three-parameter model, b parameters were estimated well by the two procedures. However, LOGIST produced more-accurate estimates than BILOG except when there were 20 items and 250 examinees. The slight superiority of LOGIST may be due to the fact that the

TABLE 3

Factors and Levels in Swaminathan and Gifford (1987)

Factors	Levels
Number of Items	20, 40, 60
Number of Examinees	250, 500, 1000
Parameter Distributions[a]	
θ	Uniform and Standard Normal
	(-1.73, 1.73)[b]
a	Uniform
	(0.6, 1.9)
b	Uniform and Standard Normal
	(-1.73, 1.73)
c	Uniform
	(0, .22)
Estimation Procedure	MML (BILOG) and JML (LOGIST)
Model	One-, Two-, and Three-parameter

[a]The distributions of the θ and the b parameters were standard normal in both of the one- and the two-parameter models. All other parameter distributions were uniform.
[b]The figures in parenthesis give the range of potential values for the parameter.

ability distribution was uniform in the simulation whereas the MML procedure assumes a normal ability distribution. Also, the uniform θ and b parameter distributions ensure better c parameter estimation by LOGIST which consequently does not have negative effects on the b and θ parameter estimation. In addition, the procedure implemented in LOGIST is not a pure JML procedure because limits are imposed on the parameter estimates.

With the a parameter, BILOG produced superior estimates with 20- and 40-item tests. When the test length was 60 items, LOGIST produced more-accurate estimates than BILOG. As noted above, in LOGIST the a estimates are constrained to a range that can be set by the user. The MML procedure, as implemented in BILOG, does not incorporate a constraint. The differential treatment of the a parameter estimates may account for the results reported by Swaminathan and Gifford. Swaminathan and Gifford stated that the better LOGIST estimates of the a parameters may be partly due to imposing a ceiling of 3.0 instead of the default of 2.0 on the LOGIST estimates of the a parameters.

Although the MML procedure does not produce ability estimates, BILOG has the facility to use the MML item parameter estimates in a maximum likelihood (ML) procedure for ability estimation. For each examinee, this procedure maximizes the likelihood given in equation (4). In these

maximizations, the a, b, and c parameters are replaced by the MML item parameter estimates. A similar procedure is used in the JML estimation of the ability parameter. In the final iteration for ability estimation, the JML is equivalent to maximizing, for each examinee, the likelihood given by equation (4). In these maximizations, the a, b, and c parameters are replaced by the JML item parameter estimates. Thus the ML ability estimates in BILOG and the JML ability estimates in LOGIST are based on the same procedure.

Differences in the ability estimates produced by LOGIST and by the ML procedure in BILOG are primarily due to utilizing JML item parameter estimates in the former program and MML item parameter estimates in the latter. Swaminathan and Gifford (1987) found that ability estimates were less accurate for BILOG than LOGIST. The ability estimates produced by BILOG were particularly poor on the lower end of the ability scale which may have been due to poor estimation of the c parameter. Thus LOGIST estimates of b, c, and θ were more accurate than BILOG estimates. BILOG was more accurate in estimating the a parameters. For each test length BILOG estimates converged to the true values as the number of examinees increased. LOGIST estimates converged only as both the number of items and the number of examinees increased.

When the two-parameter model was used, differences in the b and a parameter estimations were clear. BILOG estimates of b parameters were better with short tests. The two procedures yielded very similar estimates for longer tests and larger numbers of examinees. The a parameters were estimated more accurately using BILOG than LOGIST. With 60 items and 1000 examinees, the differences were negligible. Ability estimates were more accurate for LOGIST than BILOG when short tests were used. This trend is the opposite of that observed with item parameter estimation. LOGIST was less accurate at both ends of the ability continuum. Differences again became negligible when the number of items increased.

With the one-parameter logistic model and short tests, BILOG produced more-accurate estimates of the b parameters than LOGIST. BILOG and LOGIST produced equally accurate θ estimates for the one-parameter model. Thus as the number of parameters estimated decreases, the two programs produce equally accurate θ estimates. Differences become clear and estimates become less accurate with the introduction of the a parameters then the c parameters as well in the two- and the three-parameter models, respectively.

Swaminathan and Gifford (1987) recommended Bayesian procedures for the three-parameter model because both MML and JML were poor in estimating the c and a parameters. Either

could be used satisfactorily when ability estimation was of primary concern. However, BILOG's MML has the advantage of a more accurate ability estimation with data that fit the one- and two-parameter logistic models and meet the ability assumption of the MML procedure.

Qualls and Ansley (1985) compared the performance of BILOG and LOGIST using simulated data with the factors and levels presented in Table 4. Parameter estimates were compared with the true values and with each other by considering correlations and average absolute differences. Qualls and Ansley found that BILOG estimates (ML for the ability and MB for the item parameters) were almost uniformly more accurate than the LOGIST estimates. In nearly all cases, the BILOG estimates were more highly related to the true parameters than were LOGIST estimates. The disparity was most appreciable with smaller samples and/or shorter tests. BILOG did, however, require slightly more computer processing time.

Two problems were faced when BILOG default procedures for ability estimation were used. A high scoring examinee, who missed one easy item, was assigned the default lower bound on the ability estimates. An examinee missing one or two items may receive a higher ability estimate than an examinee with perfect scores. The use of the biweight robustification option in BILOG eliminated the problems

TABLE 4

Factors and Levels in Qualls and Ansley (1985)

Factors	Levels
Number of Items	10[a], 20, 30
Number of Examinees	200, 500, 1000
Parameter Distributions[b]	
θ	Standard Normal
a	Uniform
	(0.5, 2)[c]
b	Uniform
	(-2, 2)
c	Uniform
Estimation Procedure	ML and MB (BILOG) and JML (LOGIST)

[a]For data sets with 10 items, the simulation procedure was altered by selecting b values centered at -0.9 rather than at zero since both programs had great difficulty in deriving reasonable estimates with b values centered at zero.

[b]The data were generated to fit the three-parameter logistic model with a common c parameter.

[c]The parenthetical figures are the ranges of potential values for the parameters.

mentioned above on all but two data sets. In these two data sets the ability parameters were unestimable and the option resulted in aborted runs. One of the inconsistencies in the results of this study appeared in the ability estimates when 30-item tests were used. The LOGIST ability estimates were slightly more similar to the true parameters than were the BILOG estimates.

The generality of these findings is limited by several factors. Most importantly the abilities were simulated from a standard normal distribution, the default ability distribution assumption made in MML and MB as implemented in BILOG. In addition, the simulated tests had uniform distributions of the a and the b parameters and the c parameter was fixed at 0.2 , which are probably not typically found with real tests. Estimated abilities were not provided for all simulees (especially very high and very low-scoring examinees). Thus they contain error variance not found in the θ parameters and there is no guarantee that the estimated and the true values were actually on the same metric. Comparisons were limited to correlations and average absolute differences between true and estimated values. Also it was not stated in the paper which options were used in LOGIST and BILOG. For BILOG the options chosen could have a substantial effect on results.

The JML, the MB, and the ML Procedures

Yen (1987) compared the parameter estimation using BILOG and LOGIST. Simulated response vectors were generated using the three-parameter logistic model. The factors and the levels of the design are shown in Table 5. To create a test of moderate difficulty, 20 values of b parameters were set at -1.46, -1.25, -1.04, -0.83, -0.66, -0.59, -0.52, -0.45, -0.38, -0.31, -0.24, -0.17, -0.10, -0.03, -0.04, -0.11, 0.33, 0.53, 0.74, and 0.95. To create an easy test, 0.5 was subtracted from these b values. A difficult test was created by adding 0.5 to the moderate b values. A 40-item test was created at each difficulty level by duplicating the original 20 items. Three Reading vocabulary (RV) tests were created by using the estimated parameters of the first 10, the first 20, and the first 40 items from a reading vocabulary subtest of the Comprehensive Tests of Basic Skills, Form U(CTB/McGraw-Hill, 1982). Details of this data generation are presented in Yen (1984).

The non-normal ability distributions were obtained as mixtures of two normal distributions. The mixtures were varied to produce negatively-skewed (NS), positively-skewed (PS), and approximately symmetric but platykurtic (PK) distributions of the true ability values. The three levels of ability distributions were only completely crossed with the two levels of RV test lengths. The variations on item

TABLE 5

Factors and Levels in Yen (1985)

Factors	Levels
Number of Items	20, 40
Parameter Distributions[a]	
θ	Standard Normal Positiv Negatively Skewed (-0.4, -0.1) Platykurtic (0.1, -0.4)
a	Degenerate, all a=1.0
b	Negatively-Skewed
c	Degenarate, all c=.2
Estimation Procedure	MB, EAP and ML (BILOG) and JML (LOGIST)

[a]The three-parameter model and a 1000 examinees were used.
[b]These figures are the skewness and kurtosis of θ distributions.

difficulty were only used when the ability parameter was normal.

All default options were used in running both LOGIST and BILOG with the following exceptions. The number of answer choices were set at 4 in the two programs. An empirical prior was used for the item and ability parameter estimation of BILOG. The default ML estimates of θ parameters were also compared to the Bayesian estimates.

BILOG was more accurate than LOGIST in almost every case except for some conditions that might be related to ability distribution and will be mentioned in the distribution section of this chapter.

Yen (1987) compared some of the BILOG options for ability parameter estimation: the ML procedure and the expected a posteriori (EAP) using an empirical prior. Yen reported that these different options produced item parameter estimates that were very similar but not identical due to the slightly different numerical approximations to the ML functions that were employed. The accuracy increased with the increase of test length as indicated by the MSD (mean square deviation) and the correlation of the true and estimated values. The ability estimates of LOGIST had correlations with the true values similar to the ML ability estimates of BILOG. However, LOGIST's abilities were less biased than ML abilities as indicated by the MSDs. This

similarity conforms to previous findings (e.g., Swaminathan & Gifford, 1987).

The EAP ability estimates were more highly correlated and less biased than both the ML and the LOGIST ability estimates. The abilities estimated by the EAP procedure displayed lower RMSD (the square root of MSD) but higher local bias (the absolute mean difference averaged over Q, the cells of simulees with similar θs, divided by the standard deviation of the true values) than those based on ML procedure. This difference in local bias gives more information about accuracy of the estimate at its different levels and indicates the importance of this or similar comparison measures. A similar approach was used in this dissertation to give some information about bias in estimation of item and ability parameters at different levels of these parameters.

<u>The JML and the JB Procedures</u>

Most of the research on Bayesian estimation procedures has focused on the JB estimation procedure due to Swaminathan and Gifford. Swaminathan and Gifford (1982, 1985, 1986) have compared the JB and the JML estimation procedures using the one-, two-, and three-parameter logistic models. Tables 6, 7, and 8 present the designs used to study estimation in the one-, two-, and three-parameter models respectively.

TABLE 6

Factors and Levels in Swaminathan and Gifford (1982)

Factors	Levels
Number of Items	15, 25, 50
Number of Examinees	50, 75
True (and prior) Parameter Distributions[a]	
θ	Uniform[b], (Standard Normal)
b	Uniform[b], (Standard Normal)
Estimation Procedure	JML (LOGIST) and JB

[a]The one-parameter model was used.
[b]The mean of this distribution is zero and the standard deviation is one.

TABLE 7

Factors and Levels in Swaminathan and Gifford (1985)

Factors	Levels
Number of Items	15, 25, 35
Number of Examinees	50, 100, 200, 500
True (and prior) Parameter Distributions[a]	
θ	Standard Normal , (Uniform)
a	Uniform[b] , (Gamma)
b	Standard Normal , (Uniform)
Estimation Procedure	JML (LOGIST) and JB

[a]The two-parameter model was used.
[b]The mean of this distribution is zero and the standard deviation is one.

TABLE 8

Factors and Levels in Swaminathan and Gifford (1986)

Factors	Levels
Number of Items	25, 35
Number of Examinees	100, 200, 400
True (and prior) Parameter Distributions[a]	
θ	Standard Normal, (Uniform)
a	Uniform[b], (Gamma)
b	Standard Normal, (Uniform)
c	Uniform[b] , (Beta)
Estimation Procedure	JML (LOGIST) and JB

[a]The three-parameter model was used.
[b]The mean of this distribution is zero and the standard deviation is one.

An important characteristic of these studies is that the simulated a, b, c, and θ parameter distributions did not match the assumed prior distributions in the JB procedure. This permits some insight into the robustness of the JB procedure. In studying the one-parameter model, the priors for the θ and b parameters were standard normal distributions but the parameters were generated as samples from a uniform distribution with zero mean and unit variance. In the two- and three-parameter logistic model, θ and b parameters were assumed to be uniformly distributed but they were generated from a standard normal distribution. In the two- and three-parameter logistic models, the prior for the a parameters is a gamma distribution but the a parameters were generated from a uniform distribution. Similarly, the prior for the c parameter was a beta distribution but the c parameters were generated from a uniform distribution.

For the one-parameter logistic model, the correlations of the true and estimated ability values were equal for the JML and the JB procedures. For small values of sample size and test length, the JB difficulty estimates correlated slightly better with true difficulty values than did the JML estimates. In terms of the MSD, the JB was clearly superior to the JML procedure, particularly in the estimation of ability. The JB procedure showed the greatest advantage with small sample size and short tests.

In their study of estimation in the one-parameter model, Swaminathan and Gifford (1982) used a hierarchical JB procedure. They assumed diffuse hyperpriors for the mean of the standard normal priors for the ability and difficulty. The hyperpriors for the variances of the ability and difficulty priors were inverse chi-square distributions. In these distributions the degree of freedom was set at 10 but the scale parameter was permitted to vary at (5, 8, 15, 25). Thus , Swaminathan and Gifford provide some information about the impact of various degrees of misspecifying a hyperprior on the estimation accuracy.

It was found that hyperpriors do not seem to affect the correlation between the true values and the estimates. The MSDs indicated that the accuracy of estimation was affected to some degree, especially in small sample sizes, by the prior beliefs. This trend is more evident for the ability estimation than for the estimation of item parameters. As the degrees of freedom of the inverse chi-square distribution increase, it becomes more concentrated, the estimates regress towards the mean and higher MSDs were produced. Minimal bias was found for the inverse chi-square distribution with degrees of freedom between 5 and 15. Decreasing the scale parameter for the inverse chi-square prior distribution has the same effect as increasing the degrees of freedom. Swaminathan and Gifford used, in

conjunction with degrees of freedom between 5 and 10, a scale parameter of 10; this specification worked well.

In the two- and the three-parameter logistic models the JB estimates were superior to the JML estimates of the a, b, and θ parameters, with respect to correlations with the true values as well as the MSDs. The JB procedure consistently produced smaller MSDs than the JML procedure. This may be due to the fact that discrimination parameter is more accurately estimated by the JB procedure yielding more accurate estimates of b and θ parameters. As the number of items and examinees increase, the two procedures yield similar results, a trend that is more evident when the number of items increases. This result was expected because, in large samples, the likelihood dominates the prior and the Bayesian estimates are indistinguishable from the JML estimates (Swaminathan & Gifford, 1985, 1986). The true chance-level values correlated better with the JB estimates than with the JML estimates and had smaller MSDs, particularly with 25 items and 100 examinees.

Thus the JB procedure produced better estimates than those produced by the JML procedure particularly in the case of the a and the c parameter estimation when both sample size and test length were small. The two procedures converged to the true values as the sample size and more evidently the test length increased.

Vale and Gialluca (1988) compared four methods of item calibration. These were heuristic transformations of the traditional item statistics (Jensema, 1976), ANCILLES-X (a new verion of ANCILLES), LOGIST, and ASCAL. A 25-item test of general science was administered to 2000 examinees and the responses were analyzed by ASCAL to obtain estimates of the three-parameter logistic model. The resulting estimates were used as the true values of two data sets: Test1 and Test2. The item parameters were used twice to obtain a 50-item test (Test1) with a restricted range of difficulty. Test2, a test with wider range of difficulty, was obtained from Test1 by multiplying b parameters by 2. A 57-item test of shop knowledge was administered to 200 examinees and the responses were calibrated by ASCAL. The resulting estimates were used as the true values of Test3 which was more difficult and less discriminating than Test2. Thus Test2 was more difficult than Test1; Test3 was developed as part of an adaptive test item pool.

The θ parameters were sampled from a standard normal distribution. Smaller samples of 500 and 1000 examinees were drawn from the 2000 examinees of each of the three tests. LOGIST was used only for the data sets with 2000 examinees. Other programs were used with the three sample sizes.

The three tests showed high correlations and low RMSDs of the b and a estimates with their corresponding true values. The a estimates were best correlated with their true values in Test3 and poorest in Test2. Thus the more difficult test (Test2) produced less accurate a estimates. Test2 also had the largest RMSDs of all item parameter estimates. Therefore, the wide range of difficulty produced an adverse effect on accuracy of estimation in all item parameters. This result was previously indicated by Swaminathan and Gifford (1982) with respect to the JML estimates of the b parameters.

ASCAL consistently produced parameter estimates with the lowest RMSDs and the highest correlations of all the other calibration procedures. LOGIST invariably produced estimates of nearly equivalent quality to those of ASCAL except for the c parameter estimates which were not as good. The heuristic approximation method was the poorest, particularly at smaller sample sizes.

One problem with this study was that the true values were taken from ASCAL estimates of a science test administered to 2000 examinees. This problem might have biased the results in favor of ASCAL. This bias may be only slight because JB procedures are generally known for their superiority over the JML procedure. Therefore, using the

estimates of one program as the true values is avoided in this dissertation to ensure fair comparisons.

Summary of Sample Size and Test Length

The JML procedure as implemented by LOGIST was found to be generally superior to Urry's procedure (e.g., Ree, 1979; Swaminathan & Gifford, 1983). The JML was also found to be superior to the heurisitic approximation and ANCILLES-X (Vale & Gialluca, 1988). Swaminathan & Gifford (1987) compared the MML and the JML procedures as implemented in BILOG and LOGIST. They concluded that the MML (or ML for θ) procedure is generally superior to the JML procedure in estimating a, b, and θ parameters of the one- and two-parameter logistic models, particularly when small sample size and/or test lengths were used. For the three-parameter model LOGIST was superior in estimating b, c, and θ parameters, whereas BILOG was superior in estimating the a parameters.

Superiority of LOGIST estimates of θ parameters was because in LOGIST the a parameters were constrained to a reasonable range, the unestimable c parameters were set to a common value, and the program works better with the uniform θ used in the study. The ML estimates of θ are based on the item parameter estimates of MML which are not constrained. The a estimates greater than 4.0 were excluded upon the calculation of the MSDs for both the LOGIST and the BILOG

estimates; however, these excluded values were more in number for LOGIST than they were for BILOG estimates of the a parameters.

Yen (1987) used a broad range of generated data but a limited sample size of 1000 examinees. Using BILOG she employed the MB procedure, for item parameter estimation, and the EAP as well as the ML procedures for ability estimation. She compared these estimates of BILOG with those of LOGIST under the three-parameter model. The θ estimates of EAP were found to be better than both the ML and the JML estimates. Her study was limited only to 20- and 40-item tests with 1000 examinees. Convergence to the true values were only investigated over the increase in test length from 20 to 40 items. In spite of these limitations BILOG was not completely superior to LOGIST. The superiority of LOGIST in some cases might be attributed to the choice of generated values used in the study. Another reason for this superiority was the approach of handling extreme estimates in the two programs. BILOG pulls extreme values toward the center of the prior distribution for the item parameters, whereas that center differs a little from where it would have been without the use of the priors. In LOGIST, upper and lower bounds are placed on a and c parameter estimates to prevent them from drifting to extreme

values. This approach might have caused the superiority of LOGIST in some cases.

Qualls and Ansley (1985) used a limited range of generated data, with various levels of test lengths and sample sizes, under the three-parameter model. One important conclusion was when the ML θ estimation was used the biweight robustification eliminated the problem of assigning the lower bound ability to high scoring examinees who missed an easy item. Thus ability estimation by the ML improved over the JML of LOGIST probably because of ability robustification.

Swaminathan and Gifford (1982, 1985, & 1986), compared the JB and JML procedures. These simulation studies have shown that the JB estimates are superior to the JML estimates because they do not drift out of range, and are more accurate, even when the prior distributions are not the same as the distributions of the generated parameters. The JB estimates of ASCAL were also found to be better than LOGIST estimates (Vale & Gialluca, 1985, & 1988). The JB of ASCAL does not provide estimates of the ability parameters, is only available for micro-computers, and takes a long time running large samples and/or long tests. The JB of Swaminathan and Gifford is not currently available for general distribution.

Thus it can be concluded that the most important and available procedures for comparisons were the JML of LOGIST and the MB and MML of BILOG. Precautions were taken on generating the data so that these data were reasonable for both of the programs. For example both small and large sample sizes and test lengths were used in the comparison. The JML only converges as both the number of items and the number of examinees increase. The MML and the MB estimates were found to converge to the true values as the number of items and/or the number of examinees increase. Thus the small and the large sample-size and test-length combinations were found to be more important and more reasonable than other combinations. With these combinations, convergence of parameter estimates to their true values can be proven or disproven with the increase of sample size and/or test length. The number of items and the number of examinees chosen in this dissertation were designated as small and large in accordance with some of the aforementioned studies (e.g., Swaminathan & Gifford, 1987).

IRT Parameter Distribution and Estimation Procedures

Earlier it was pointed out that the JML procedure does not incorporate any assumptions about the distributions of item or ability parameters. The MML procedure requires an assumption about the distribution of ability. The JB and MB

procedures require assumptions about the priors of both item and ability parameter distributions. There is a small body of literature relevant to the impact of violations of the distributional assumptions of the latter three procedures and the degree to which the efficacy of the JML procedure is affected by variations in IRT parameter distributions. This literature is reviewed in this section.

Urry's Method and the JML Procedure

In the Swaminathan and Gifford (1983) study described in the preceding section, three ability distributions were simulated: normal (N), negatively skewed (NS), and uniform (U). The effect of various ability distributions on accuracy of estimation was most obvious in the case of the a parameter estimation and least obvious in the case of ability estimation. For the JML procedure, the highest correlations between true values of the a parameters and the JML estimates were obtained when the uniform ability distribution was used. In general, the NS ability had a negative impact on JML estimation of the a parameters. For example, with 20 items and 1000 examinees, the a estimates had a correlation of .52 with the true parameter when θ was skewed to the left. The correlations increased to .56 and .76 when the θ parameters were uniform and normal, respectively. Except for a few cases, the a estimates with normal θ had higher correlations than those with uniform θ.

For longer tests the correlations improved across the three ability distributions.

The estimates of the a parameter were negatively correlated with the true values of short tests when the ability distribution was negatively skewed. The estimation of b and θ did not seem to be affected by the ability distribution. For example, with 20 items and 1000 examinees, θ estimates had correlations of .88, .89, and .91 with the normal, uniform, and skewed θ parameters respectively. LOGIST underestimated the c parameter; however, the estimates were most reasonable with normal ability distribution. The LOGIST estimates of ability resulting from a skewed distribution of ability were as good as, and in some cases better than, the estimates obtained with a normal distribution. Swaminathan and Gifford (1983) indicated that although the uniform distribution had a larger chi-square value (a measure of deviation from normality) than the skewed distribution, the results obtained with uniform distribution were similar to those obtained with normal distribution. It is then not departures from normality but departure from symmetry and the unavailability of examinees in the lower tail of the ability distribution that affected estimation procedure.

One strength of this study was that the convergence of the JML estimates of the three-parameter model was

investigated under three θ distributions. It was found that the JML estimates converge to their true values as both number of examinees and test length increased even when θ distribution was varied. Convergence of MML or MB was not investigated for various θ distributions, not in this study or in any other study reviewed in this report.

One weakness in this study was the use of correlation coefficients as the only measure of comparison for a, b, and θ estimates with their true values. Means and standard deviations were used in the comparison of c estimates because the generated c was fixed at 0.25 for all items. Total bias as well as bias at several levels of estimates is a very important measure of comparison that indicates the potential of some procedures to estimate upper and lower limits of parameters. Also the coefficients of skewness and kurtosis were not reported.

Ree (1979) conducted a simulation to compare item and ability parameter estimates produced by three computer programs: LOGIST, ANCILLES, and OGIVIA. ANCILLES is a newer version than OGIVIA which is suitable for approximate estimation of item and ability parameters in the three-parameter model. The two programs are based on the procedures presented by Urry, (1977). These procedures provide good estimates for large numbers of examinees and items. Three ability distributions were employed: normal,

TABLE 9

Factors and Levels in Ree (1979)

Factors	Levels	
Number of Items	80	
Number of Examinees	2000	
Parameter Distributions[a]		
θ	Standard Normal	(-2.5, 2.4975)
	Positively Skewed[b]	(-0.506, 2.379)
	Uniform	(0, -1.2)
a		(0.653, 1.6136)
b		(-1.653, 1.9745)
c		(0.0872, 0.3479)
Models	Three-parameter	

[a]The figures in parentheses give the range of potential values for the parameter.

[b]The skewed ability was generated by selecting 2000 θs from 3000 cases of a unit normal distribution. A cutting score was set to yield the upper two-thirds of the population.

positively skewed, and uniform. Estimates of skewness and kurtosis were (0, -1.2009), (0.6359, -0.2698), and (-0.0050, 0.1144) for the uniform, positively skewed, and normal distributions; respectively. A sample size of 2000 simulees for each ability distribution was used to simulate 80 five-option multiple-choice test questions.

In spite of the large sample size and test length used in this study, some effects of θ distributions on accuracy of estimation were found. One possible reason was the use of a different a or c parameter for each item as opposed to using a fixed value for all items (e.g., Yen, 1987). Another reason was the fact that skewness of θ was almost twice as much as the one used by Yen and showed no effect on the accuracy of estimation. Thus the varying of a and c parameters and the doubling of skewness of θ have caused an effect on the JML estimation accuracy.

For normal, positively skewed, and uniform ability distributions the correlations between true cs and the cs estimated by LOGIST were .379, .233, and .557. The correlations for the as, the bs, and the θs were (.565, .827, .895), (.447, .975, .978), and (.943, .965, .974) respectively. Thus similar trends, to those in Swaminathan and Gifford, were found for the a, b, and c parameter estimates with the least accurate estimation indicated when the skewed ability distribution was used. The accuracy of θ

parameter estimation was not as strongly affected by the variation of its distribution. This small effect on θ estimation is in agreement with the results reported by Swaminathan and Gifford (1983).

The accuracy of a and c parameter estimation was best when uniform θ was used. In spite of using large sample size and test length and using correlations as the only measure to compare accuracy, Ree concluded that the selection of an item calibration program should be dependent on the distribution of the ability in the calibration sample and on the computer resources available. By using measures of bias and small and large sample sizes and test lengths, one may find clearer differences and may identify convergence to the true values.

Wingersky and Lord (1984) found that using a rectangular θ distribution yielded smaller standard errors of the JML item parameter estimates than did doubling the number of items under a bell-shaped θ distribution. When as and cs were low ($b-2/a < 1$), the standard errors obtained with the rectangular θ distribution were as low as those obtained with normal θ and quadruple the number of examinees. Results supporting this finding were reported by by Ree (1979) and by Swaminathan and Gifford (1983).

The JML, the ML, and the MB Procedures

Yen (1987) compared the ML ability estimates and the MB item parameter estimates produced by BILOG to the JML estimates of LOGIST for easy, moderate, and difficult tests; when the true abilities were normal and non-normal (i.e., NS, PK, and PS). The variation of the ability distribution had little effect on the accuracy of the JML difficulty estimates. The correlations of the true bs with their LOGIST or BILOG estimates were almost identical across the four ability distributions. The MSDs of the true bs and their JML estimated values were (0.18, 0.26, 0.19, 0.20) and (0.12, 0.17, 0.15, 0.16) for normal, NS, PK, and PS distributions of the 20 item and 40 item tests, respectively. For the MB, the MSDs were (0.11, 0.13, 0.12, 0.09) and (0.11, 0.11, 0.12, 0.13) for the four ability distributions. With the 20 item test, the JML was least accurate when θ was negatively skewed and most accurate with normal θ distribution. Accuracy of the JML b estimates improved, as indicated by the MSDs, when the number of items increased to 40. The MSDs of the b estimates by the MB procedure were very similar across θ parameter distributions and across test lengths.

The correlations for θ were almost identical across the four distributions in the two programs. The corresponding values of local bias showed some variation especially with

the JML estimates of the 20-item tests. The values of local bias of the four distributions were (0.06, 0.07, 0.04, 0.05) and (0.05, 0.02, 0.03, 0.03) for the 20- and the 40-item tests respectively. Local bias values were the smallest for the PK ability distribution. The values of local bias decreased for the four ability distributions when the number of items increased to 40.

Yen (1987) also compared two ability estimation procedures implemented in BILOG: the ML and the EAP procedures. The correlations of the estimated and true values of θ were more consistent across the four ability distributions for the EAP than for the ML procedure. The EAP procedure was hierarchical: The means and the standard deviations of the prior distribution are updated at each iteration of the estimation procedure. The updating may have caused the consistency of estimation across ability distributions. The ML estimates, however, were also consistent across various ability distributions. Again no drastic changes were shown by the RMSDs or the correlations.

When the a parameters were estimated, the correlations with the true values were a little more varied. These correlations were (.82, .97, .93, .95) and (.90, .90, .90, .94) for the JML estimates of the a parameters in the 20 item and 40 item tests for the MB estimates correlations were (.92, .93, .94, .93) and (.88, .92, .91, .94) for the

20 item and 40 item tests across the four distributions. No drastic changes were shown by the MSDs except with normal θ parameter. For 20-item tests the accuracy of a estimates by JML were shown to be the poorest when normal θ parameter was used. The MSDs of the JML were (0.73, 0.16, 0.21, 0.16) and (0.23, 0.18, 0.16, 0.16) for the 20-item and 40-item tests respectively. Thus the MSDs of the a estimates were improved for the normal θ parameter when the test length increased to 40.

Comparing the accuracy of the two programs, Yen found that in almost every case the item parameters produced by BILOG were more accurate than those produced by LOGIST. LOGIST produced better estimates in some cases. In the 40-item tests LOGIST produced better estimates for a and θ, for c and θ, for a and b, and for a and c when θ was normal, PK, PS, and NS, respectively. The θ estimates mentioned above were only better than the ML estimates, not the MB estimates, of BILOG. These differences between the two program indicate that LOGIST might be better than BILOG with certain ability distributions in the three-parameter logistic model and under similar data and model characteristics of Yen's study. In spite of these slight differences, the study of the effect of ability distributions on accuracy of estimation is motivated by this

study. The reason is the limitations that might have caused these slight differences to occur.

This study was limited by several factors. The variation in the distribution of the θ had only a slight effect on the accuracy of the procedures investigated by Yen because she only used a large number of examinees (1000), a slightly skewed and platykurtic θ (0.1, 0.4), and a constant value for each of c and a parameters for all items. Thus the large number of examinees, the slight skewness, the slight kurtosis and the constant a and c, simplified the iterative solution so that a slight effect of θ distribution was found on accuracy. Larger coefficients of skewness and/or kurtosis of θ distributions were thus motivated by Yen's study, when both a and c parameters were varied and when large and small numbers of examinees were combined with large and small numbers of items. In Yen's study the number of examinees was not varied across the θ distributions and the effect of only slightly skewed and kurtic θ distributions on accuracy was compared.

Swaminathan and Gifford (1983) reported the prevalence of the effects of varying ability distributions with small sample size and test length. The effects become negligible and the estimates converge to the true values, with the increase of the number of both items and examinees. Swaminathan and Gifford (1983) and Ree (1979) used larger

coefficients of skewness and kurtosis than those used by Yen. They did not investigate the effect of these larger coefficients on accuracy of estimation using BILOG.

Rigdon and Tsutakawa (1983) compared the MML, MLF, and the CMLF of the one-parameter logistic model. The ability parameters were chosen randomly from a standard normal distribution to represent 50 and 200 hypothetical examinees. Three sets of 50 difficulty parameters were chosen non-randomly as the 1, 3,..., 99 percent points, of three distributions. The three distributions were: normal (concentrated near the average ability), uniform over the range $-3\frac{1}{2}, 3\frac{1}{2}$ (spread out uniformly), and U-shaped difficulty (sparsely near the average ability). Six response matrices were then generated from the ability and difficulty parameters.

The MSDs of the MML were larger than those of either the MLF or the CMLF. The MSDs of the MLF and the CMLF were comparable. They both were 9% lower than for θ estimates and 19% lower for b estimates.

The MSDs of the θ estimates did not differ across the three distributions of the b parameters. However, a slight difference was found in the MSDs of the b estimates across the three distributions of b parameters, particularly, when sample sizes were small.

Rigdon and Tsutakawa (1987) also compared the MLF, the CMLR, and the CMLU under the one-parameter logistic model. They used the same data used in their 1983 study. In most cases the MSDs of the three procedures were close. The MSDs of the CMLR estimates of the b parameters were considerably less than those of the corresponding CMLU and the MLF estimates. The CMLR was thus found to be robust with respect to the normal assumption violation of the b parameter. The CMLR was recommended by Rigdon and Tsutakawa for cases with few examinees and limited information about the item response curves. The differences in MSDs of θ estimates across the three distributions of the b parameters were not substantial. The lowest MSDs were those of the uniformly distributed b parameters. In spite of the fact that the CMLR was robust against variations in b parameter distributions and with small sample size, it was not used in this dissertation. The CMLR procedure was not used in this study because the program that implements it is not available for distribution and the procedure is limited to the one- and the two-parameter logistic model.

Summary of IRT Parameter Distribution

Swaminathan and Gifford (1983) varied the θ parameter distributions and found it had a little effect on JML estimation of the b and θ parameters but did affect estimation of a and c estimates of these parameters. The a

and c estimates were less accurate with the negatively skewed θ than with the uniform or normal θ. The uniform ability distribution produced more accurate a and c estimates than the normal ability distribution. Ree (1979) also found the poorest item parameter estimates with the positively skewed θ parameter distribution and the best item parameter estimates with the uniform ability distribution. The two studies did not include the MML nor the MB procedures in the comparison. They both reported differences in accuracy of estimation due to θ parameter distribution and provided some insight with respect to the importance of varying sample size and test length in addition to varying the θ parameter distribution. The JB procedures were also found to be superior to the JML of LOGIST (Swaminathan & Gifford, 1982, 1985, & 1986) because they do not drift out of range. They were more accurate even when the prior distributions were different from the generated values.

All of the preceding studies used only the correlation of estimates with true values except for Yen who used the MSDs as well. It is important to investigate bias, MSD, and variance of these estimates compared to their true values at several levels of both item and ability estimates. None of the preceding studies provided such comparative measures under several levels of the estimates. Swaminathan and Gifford

(1987) reported differences of practical interests at several estimate levels; however, they used only uniform distributions, which worked well with LOGIST, in the three-parameter model. Thus it is important to investigate differences in estimate accuracy across several distributions and to include ability distributions that do not favor one program over another.

Another common factor was varying sample size and test length to show convergence across the ability distribution. The two studies that used large and small numbers of items and numbers of examinees were the study by Swaminathan and Gifford (1983) and that of Wingersky and Lord (1985). The two studies did not investigate the procedures of BILOG. In the latter study LOGIST estimates were used as the true values.

Common to all of the above studies, with the exception of Yen's, was the absence of comparing the effects of ability distributions on estimation accuracy of MB, MML, and JML. Yen (1987) varied the θ parameter distributions and included the JML, MB, ML, and EAP procedures. She kept both the a, the c, and the number of examinees at constant values. The θ parameter distributions used were slightly kurtic and slightly skewed. Therefore varying the distribution of θ had only a slight effect on the accuracy of the procedures investigated by Yen.

The distributions of the b parameters were also varied by Yen (1987) and by Rigdon and Tsutakawa (1987). Rigdon and Tsutakawa recommended the CMLR for small sample size and non-normal b parameter distributions. Because the CMLR program is not available publically and is restricted to the one- and the two-parameter models, the CMLR was not used in this study. The MB procedure of BILOG was used instead.

Among the non-normal ability distributions used in the literature are the uniform and beta distribution used by Swaminathan and Gifford (1983), the truncated normal distribution used by Ree (1979), and the skewed and the platykurtic distributions used by Yen (1987). The beta and the truncated normal distributions were selected for the present study because these are realistic distributions and have had a negative impact on estimation in previous studies. The uniform distribution is unrealistic. Yen's distribution apparently did not deviate sufficiently from normality to have an effect on estimation accuracy.

CHAPTER III
METHODOLOGY

Design of the Study

To accomplish the purpose of the study a simulation approach was used. The conditions varied were the distribution of the ability parameter, the number of items, and the number of examinees. These conditions were combined to provide various cells of the design. For each cell of the design, the data were replicated 10 times and each replication was calibrated by the JML of LOGIST and the MML and the MB of BILOG. Details of the levels of these conditions and methods of generating the data are described in the following sections.

Sample Size and Test Length

In the present study, sample size and test length each had two levels. These were the small and the large levels used by Swaminathan and Gifford (1987). The numbers of items used were 20 and 60, whereas the number of examinees were 250 and 1000. The importance of including small and large sample sizes and test lengths was to investigate convergence of the three calibration procedures when the ability distribution was not normal. For example

Swaminathan and Gifford (1983) found that with the increase of sample size and test length, the effect of varying the ability distribution on accuracy of estimation becomes negligible.

Parameter Distributions

For each combination of sample size and test length, each of three ability distributions was employed. In the first combination the distributions were standard normal for θ and b, lognormal for a, and beta for c. The MB procedure in BILOG can be implemented with these distributions as the assumed distributions. No assumptions are made about the parameter distributions in LOGIST. In the MML procedure an assumption is made about the distribution of θ, but not about the distributions of a, b, or c. The default assumption in BILOG is that θ is standard normal. Consequently, the first distribution combination also meets the assumption of the MML procedure as implemented in BILOG. The standard normal θ is often used by LOGIST users (e.g, Wingersky & Lord, 1984). This permits a fair comparison of the MB, the JML and MML procedures under conditions that are reasonable for the three procedures.

The effect of violating the assumption of normal ability was investigated by varying the distribution of the ability parameter in the second and the third combination. In the second combination the ability distribution was generated

from a beta distribution with parameters 5 and 1.5. In the third combination the ability parameter were generated by truncating a standard normal distribution at a cutoff score of -0.053 by selecting values above that score. The distributions of a, b, and c parameters of the second and the third combinations were the same as in the first combination.

Parameter Intervals

In the process of selecting parameter intervals, reasonable for BILOG and LOGIST, several simulation studies were reviewed. Hattie (1984) used two ranges of θ to study several indices of unidimensionality. These were (-1, +1) and (-2, +2). Hambleton and Cook (1983) used the same two ranges for the b parameter to study the effect of sample size and test length on precision of ability estimates. They also used the two ranges (0.5, 2) and (0.6, 1.5) for the a parameters, and the value of 0.25 for the c parameter. Samejima (1986) used the ability range (-2.5, 2.5) to study the effect of using the three-parameter model for estimation when the data fits the two-parameter model. Lord (1977, 1980) used the ranges (-4, +4) and (-5, +5) for θ. Swaminathan and Gifford (1982) used the ranges (-1.73, 1.73) for θ, (-2, 2) for b, (0.6, 2) for a, and 0.25 for c. Hulin, Lissak and Drasgow (1982) used the ranges (0.3, 1.4), (-3, 3), (0.11, 0.33), and (-3, 3) for a, b, c, and ,θ

parameters, respectively. They applied a power of 1.4 to the a values to make the a distribution skewed. Ree (1979) used the ranges (0.5, 2.5), (-3, 3), (0, 0.3), and (-2.5, 2.5) for a, b, c, and θ, respectively.

Hambleton, Murray, and Williams (1983) used the range (0.4, 2) for a and the range (0, 0.25) for c, to study item misfit in applying the one-parameter model to data that fit the three-parameter model. Hambleton and Swaminathan (1985) used three ranges for b: homogeneous (0, 0), moderately wide (-1, +1) and wide (-2, +2); the range (0.6, 2) for a; the range (0.0, 0.25) for c; and two ranges for θ: (-2, +2) and (-3, +3), to study the effect of varying ranges of b parameters on accuracy of estimation. As mentioned earlier, it was found that wide ranges of b parameters, (-3, 3), had an adverse effect on the JML estimation of b (Swaminathan & Gifford, 1982) and an adverse effect on ASCAL and JML estimation of item parameters (Vale & Gialluca, 1988). Therefore a narrower range for the b parameters, such as the one used by Swaminathan and Gifford (1983) was used in this study.

The preceding investigators were mainly of LOGIST. Investigators of BILOG and LOGIST were also found to employ similar parameter ranges. Swaminathan and Gifford (1987) used the range (-1.73, +1.73) for both the θ and b parameters, (0.6, 1.9) for a, and 0.22 for c. Quals and

Ansley (1986) used the range (0.5, 2) for a and the range (-2, 2) for the b. On using a standard normal b parameter that is multiplied by 2, we also found several extreme item parameter estimates of the MML procedure.

The method of controlling ranges used in this study was to try different seeds until the desired range was obtained. This method was used in generating a, b, and c parameters to fall in the ranges (0.363, 2.478), (-2.19, 2.32), and (0.009, 0.343) respectively. The normal distribution was generated to have the range (-3.142, 3.02), the truncated normal and the beta distributions were generated to have the ranges (-1.534, 4.210) and (-3.635, 1.484) after standerdization. These were considered reasonable ranges in the terms of the articles reviewed. In this dissertation, no study was made of the parameter ranges except the variations in the ability ranges due to the variations in the ability distributions.

IRT Model

Mislevy and Stocking (1987) stated that most of the same estimation problems arise under all three models. They also indicated that because the one- and the two-parameter models can be expressed as special cases of the three-parameter model, any solution to the problems of the three-parameter model applies to the simpler models as well (although some solutions for the one-parameter model do not generalize to

the two- or the three-parameter models). Therefore, the three-parameter model was used in this study.

Data Generation

The data simulated in this study were generated using a data generator similar to DATAGEN (Hambleton & Rovinelli, 1973) but capable of manipulating the IRT parameter distributions as required by this study. The program can generate data for any number of examinees, any number of items, and any number of dimensions. The maximum and minimum values of each parameter can also be changed to any range, either by using special formula or by trying several seeds. The latter procedure was used in this study.

The following steps describe the data generation for the three-parameter logistic models.

1. Specify the number of items.
2. Specify the number of examinees.
3. Use a suitable seed to produce a reasonable interval for the parameter generated.
4. Generate θs from the distribution(s) of interest.
5. Standardize the distribution, when the truncated normal or the beta distribution is used.

   The mean and the standard deviation of the beta

distribution were taken from Table II of incomplete beta distributions by Pearson and Hartley (1956, p. 436). The mean and the standard deviation of the truncated distribution were derived in appendix B, and calculated using the formulae:

$\mu = -(3/2)(2\pi)^{-\frac{1}{2}}\ (e)^{-\frac{1}{2}c}$

$\sigma = (3/4\ [\ 1 + IG(c/2;\ \alpha = 3/2,\ \beta = 1)] - \mu^2\ )^{-\frac{1}{2}}$

where c is the square of the cutoff score, and IG is the integral of the incomplete gamma function with parameters c/2, α, and β. The integral was obtained from Table I of the incomplete Γ-function by Pearson and Hartley (1956, p. 2).

6. Repeat steps 3 and 4 for item parameters.
7. Calculate $P_i(\theta_j)$ using equation (1).
8. Generate a random number x_{ij} from a uniform distribution on the closed interval zero to one.
9. Generate item response u_{ij} for the three-parameter model by comparing x_{ij} with $P_i(\theta_j)$. If x_{ij} is less than or equal to $P_i(\theta_j)$, then $u_{ij} = 1$, otherwise $u_{ij} = 0$.
10. For each cell or factor combination, repeat steps 1 to 9, to obtain 10 replications for more accurate and stable results.

Method of Comparison

Indices of Comparison

After the item scores were generated, parameter estimates were then obtained using LOGIST and BILOG. Four measures of accuracy were adopted in the present study: the correlation of the estimates and the true parameter, the bias, the variance, and the MSD of the estimators. Replication of every cell in the design was essential to understand the stability of the estimation procedure. It is only through replication that the bias and the variance of an estimator can be accurately estimated.

Developing Common Metrics

Before the calculation of the bias, the variance, and MSD accuracy measures, the estimates have to be rescaled so that these estimates and true values can be compared with each other. The item response model given in (1) is indeterminate in the sense that adding a constant to both the b_i and the θ_j does not change the quantity $a_i(\theta - b_i)$. Consequently $P_i(\theta)$ does not change. That is, the choice of origin with respect to the difficulty and ability parameter is purely arbitrary. Multiplying b_i and θ_j and dividing a_i by the same constant also does not change $P_i(\theta)$. That is,

the choice of unit for measuring θ_j and b_i is purely arbitrary. The estimates produced by LOGIST and BILOG are expressed on scales that are internal to each of these programs.

Given that the true values are known, the choice of the common metric can be made so that the estimates and the parameters are comparable. Scaling can be done on either b_i or θ_j. In this study scaling was done on b_i. The equations of linear transformations are

$a_{i2}^* = a_{i2}/A$,

$b_{i2}^* = A\, b_{i2} + B$, and

$\theta_{i2}^* = A\, \theta_{i2} + B$

where A is the slope and B is the intercept.

To select a scaling procedure, current methods of calculating A and B are reviewed. Marco (1977) found A and B so that the mean and standard deviation of the transformed distribution equal to the mean and the standard deviation of the estimated item difficulties. In this method, sample moments, can be seriously affected by poorly estimated item difficulties. Stocking and Lord (1983, p. 203) stated that Cook, Eignor, and Hutton (1979) have attempted to solve this problem by restricting the range of item difficulties. Linn, Levine, Hastings and Wardrop (1980) employed weighted moments to reduce the influence of these outliers. All points with the same standard error are treated in the same

way regardless of status of their outliers. Stocking and Lord (1983, p. 203) reported that Bejar and Wingersky (1981) used a robust method that gives smaller weights to the outlying points. They treated all outliers in the same fashion regardless of their standard errors.

Stocking and Lord (1983) presented a method to overcome potential problems of the procedures by Linn et al. and those by Bejar and Wingersky. This new method gives low weights to poorly estimated parameters and to the outliers. A drawback of this method is that only the information contained in b parameters is used whereas the information in a parameters is ignored. Stocking and Lord calculated A and B by minimizing the mean square difference between two test characteristic curves (TCC). This method includes the information in a and b parameters in the calculation but it does not take into account the standard errors of estimates. The chi-square method, by Divgi (1985), is simpler and cheaper than the TCC method. It makes use of the information about sampling error which is not accounted for in the TCC method. Therefore the chi-square method was used in this dissertation.

CHAPTER IV
RESULTS

Introduction and Summary

In this study, estimation of the item and ability parameters of the three-parameter logistic model was investigated by using simulation methods. The cells of the design were defined by combinations of sample size, test length, and ability distribution. For each cell of the design, the data were replicated 10 times and each replication was calibrated by the JML of LOGIST, and the MML and the MB of BILOG. Details of the levels of these conditions and methods of generating the data were described in Chapter I.

The primary objective of the simulation was to compare the accuracy of three estimation procedures under three ability parameter distributions. The three procedures are the JML as implemented in LOGIST, and the MML and the MB as implemented in BILOG. Maximum likelihood (ML) estimation was used for estimating ability parameters when either the MB procedure or the MML procedure was used for estimating item parameters. The ML procedure will be called the ML-MML in the former condition and ML-MB in the latter. For item parameters the following accuracy indices will be reported: correlation percentiles, mean squared deviation (MSD),

squared bias, and variance. In addition, variance and bias has been depicted for all parameters by obtaining scatterplots of estimates against the true values. For ability estimates, the average MSDs were evaluated within each of nine intervals, so that the accuracy of ability estimation could be ascertained and reported.

One of the major trends in this study was that for all item parameters, sample sizes, and test lengths the MB estimation procedure produced more accurate item parameter estimates than the MML or the JML estimation procedures. Another trend was the differences in accuracy of estimation under the three ability distributions. The MB procedure produced more accurate estimates with the normal distribution than with the beta or the truncated normal ability distributions. Superiority of MB estimates was more obvious with small sample size and/or short tests. The MB and the MML estimates converged to the true values by increasing sample size or test length; the JML estimates did not. The JML estimates of θ were more accurate than the ML-MML or the ML-MB estimates for each of the three ability distributions. The ML-MML and ML-MB tended to be more accurate for the normal ability distribution than they did for the other distributions. Thus estimation accuracy was found to depend on ability distribution, sample size, test length,

and estimation procedure. The details of these major trends will be discussed in chapter V.

The results in this study are presented for sample size and test length combinations in the following order: 20 items by 250 examinees, 20 items by 1000 examinees, 60 items by 250 examinees, and 60 items by 1000 examinees. Within each combination the parameter estimates are compared in the following order: a, b, c, and θ. For each parameter estimate, comparisons are presented for the three estimation procedures within each of the three ability distributions. The correlation results are reported first, then the MSDs, and finally the bias and the variance.

Short Tests and Small Sample Sizes

For 20 items and 250 examinees, the estimation accuracy indices are reported in Tables 10, 11, 12, and 13. For each procedure-distribution combination, there are 10 correlation coefficients. For each of these 10 coefficients, the median, the upper quartile, and the lower quartile of the distribution of correlations are reported to give some idea of the variability over replications.

Accuracy of the a Parameter Estimation

An examination of the correlations in Table 10 indicated that within each ability distribution, the MB procedure had the highest median correlations, followed by the MML, and

finally the JML procedure. Similarly, the MSDs for the MB procedure were smaller than for the other two procedures. Squared bias and variance were also smaller. With the beta distribution the MML procedure had the smallest bias. In disagreement with the results for the correlations, the MSDs for the MML estimates were higher than those for the JML estimates of the a parameters within each ability distribution. The difference in MSDs for the MML and the MB is largely due to differences in variance: The bias of the two procedures was approximately the same within ability distributions. Three scatterplots are presented in Figure 1. Each is for the MML estimates and is a plot of the 10 replications for estimated a parameters against true a parameters. Moving counterclockwise from the upper left quadrant of the page, the scatterplots are for the normal, truncated normal, and beta ability distributions respectively. Similar scatterplots are presented in Figures 2 and 3 for the MB and JML procedures respectively. On each scatterplot mean estimated as are indicated by a * and a 45° agreement line is drawn for reference purposes. The reason for the larger MSDs for the MML procedure can be seen by comparing corresponding scatterplots in the three figures. Points in the scatterplots of Figure 1 are more scattered away from the agreement line than are the points in the corresponding plots in Figures 2 and 3. The JML estimates

TABLE 10

Accuracy Indices for a Parameter Estimates: 20 Items and 250 Examinees.

Estimation Procedure	Correlation Percentiles			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
MML	.51	.74	.84	0.599	0.104	0.495
MB	.80	.86	.89	0.178	0.061	0.117
JML	.33	.53	.56	0.375	0.125	0.250
Truncated θ						
MML	.45	.67	.77	0.470	0.120	0.350
MB	.69	.77	.86	0.197	0.077	0.120
JML	.48	.53	.59	0.350	0.126	0.224
Beta θ						
MML	.60	.73	.85	0.479	0.094	0.385
MB	.83	.86	.90	0.317	0.143	0.174
JML	.39	.54	.57	0.366	0.124	0.242

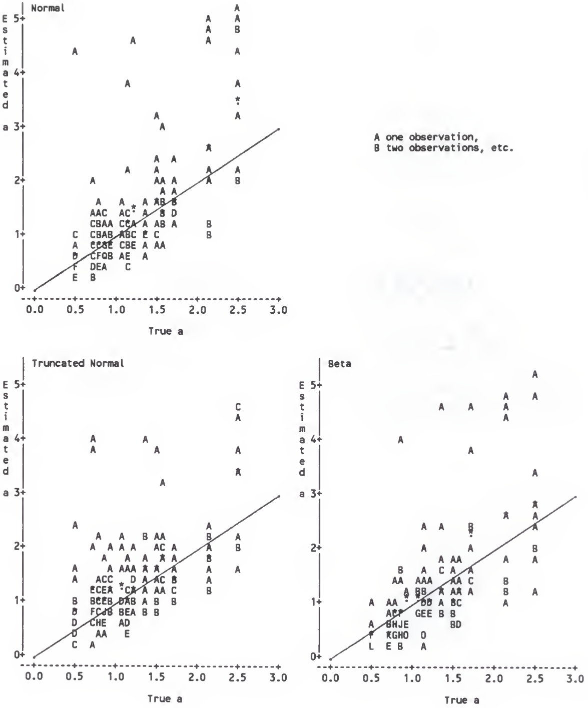


Figure 1. Scatterplots of MML Estimates of a Parameters for 20 Items and 250 Examinees.

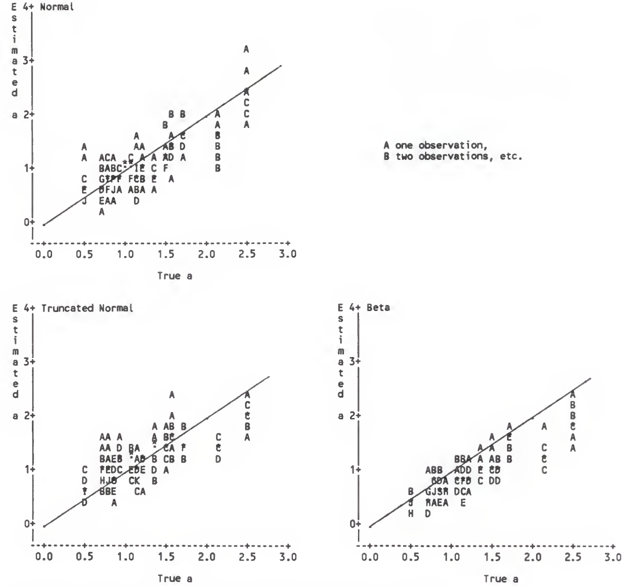


Figure 2. Scatterplots of MB Estimates of a Parameters
for 20 Items and 250 Examinees.

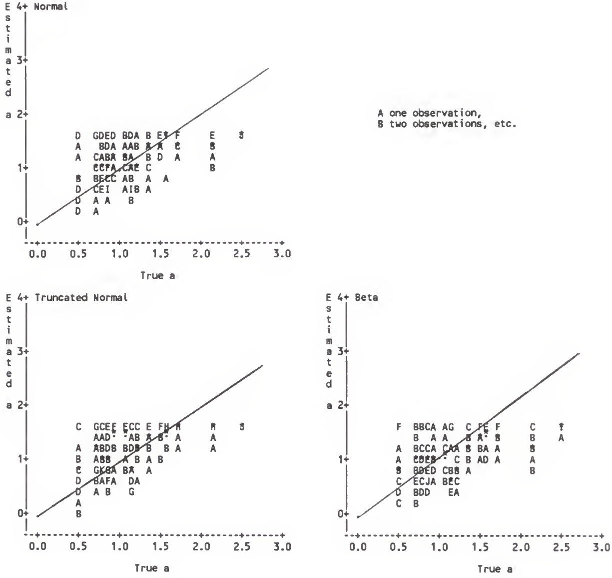


Figure 3. Scatterplots of JML Estimates of a Parameters for 20 Items and 250 Examinees.

of the a parameters, as presented in Figure 3, were more accurate than the MML estimates but less accurate than the MB estimates of the a parameters. The reason MSD is lower for the MB than for the JML can be seen by comparing Figures 2 and 3: The JML estimates appear to be more negatively biased than the MB estimates are when the true as are large and to have larger variability when the true as are small.

For the JML procedure, the median correlations for the a parameter estimates were very similar across the three ability distributions. For the MML and MB procedures, there were larger differences across the three ability distributions. For each, the lowest correlation was observed when the truncated normal distribution was used. The MSD, bias, and variance for the JML procedure were affected by the ability distribution to only a small degree. The MB procedures appeared to work more poorly with the beta distribution than with the other ability distributions. For the MML procedure, the variance decreased in moving from normal to truncated normal or beta distribution with consequent decrease in MSD. These decreases in variance are depicted in Figure 1. Comparing the plots for the normal and truncated normal ability distributions, fewer extremely large estimates are observed in the latter scatterplot. Comparing the scatterplots for the normal and the beta distributions and focusing on the lower end of the true a

scale, there seems to be a smaller degree of scatter around the agreement line for the latter.

<u>Accuracy of the b Parameter Estimation</u>

The median correlations for the b parameter estimates were highest for the MB procedure as indicated in Table 11. The JML and the MML procedures had similar median correlations. Within the normal and the truncated normal ability distribution, the MSDs indicated that the MB procedure had the most accurate estimates of the b parameters. When the ability distribution was beta, the MSDs were equal for MB and MML estimates of the b parameters. Within each ability distribution, the MSD for the JML estimates was about five or six times as high as that of the MML or the MB estimates of the b parameters. Both the squared bias and variance components of MSD are larger for the JML estimates than for the MB or MML estimates. These larger MSDs for the JML estimates, in comparison to the MB and MML estimates, of the b parameters are depicted in Figures 4, 5, and 6. Particularly, at the lower end of the difficulty scale, JML estimates are less accurate. As shown in Figure 4, the MML estimates of the b parameters were also poor at the lower end of the difficulty scale except when the ability distribution was normal.

TABLE 11

Accuracy Indices for b Parameter Estimates: 20 Items and 250 Examinees.

Estimation Procedure	Correlation Percentiles			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
MML	.89	.93	.96	0.254	0.051	0.203
MB	.93	.97	.98	0.217	0.065	0.152
JML	.91	.92	.93	1.262	0.575	0.687
Truncated θ						
MML	.90	.92	.95	0.272	0.095	0.177
MB	.96	.97	.97	0.262	0.105	0.157
JML	.89	.91	.92	1.448	0.678	0.770
Beta θ						
MML	.87	.91	.92	0.484	0.250	0.234
MB	.93	.95	.95	0.484	0.202	0.282
JML	.93	.93	.94	2.393	1.120	1.273

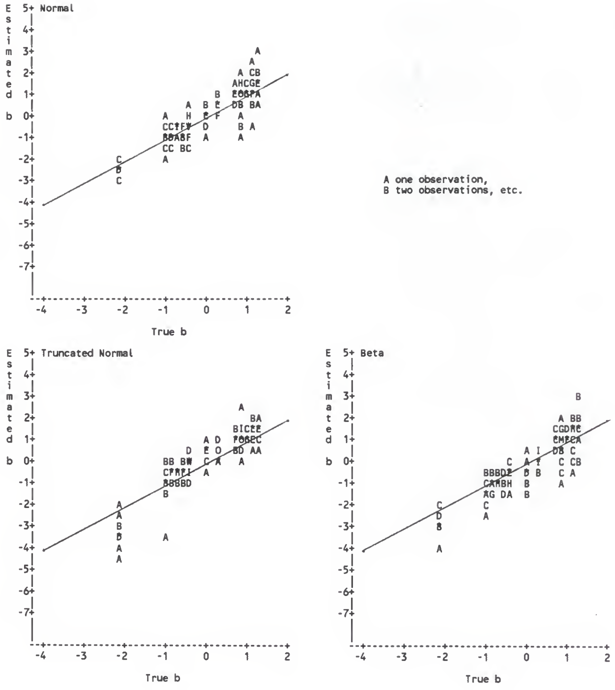


Figure 4. Scatterplots of MML Estimates of b Parameters for 20 Items and 250 Examinees.

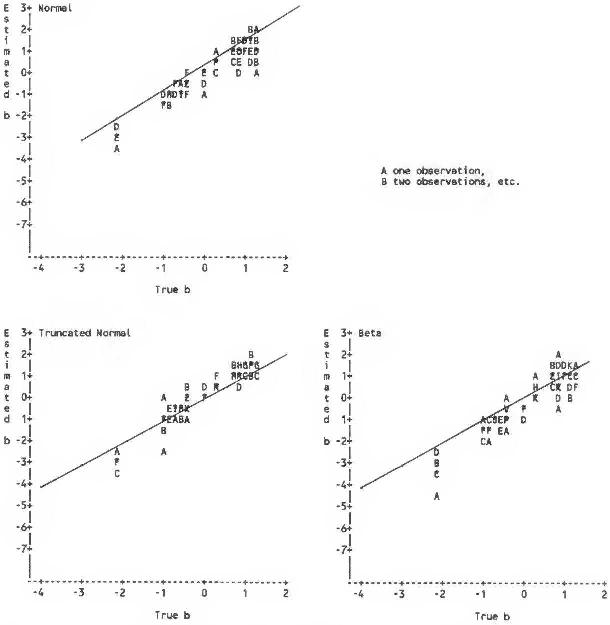


Figure 5. Scatterplots of MB Estimates of b Parameters
for 20 Items and 250 Examinees.

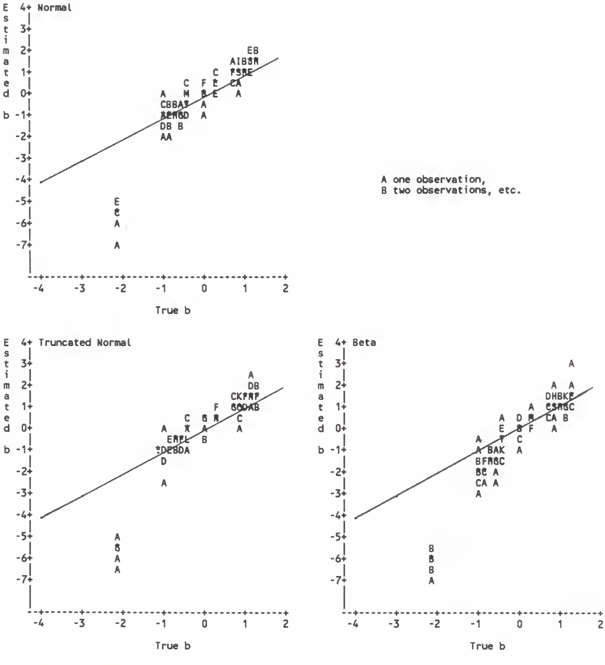


Figure 6. Scatterplots of JML Estimates of b Parameters
for 20 Items and 250 Examinees.

For each of the estimation procedures, there were appreciable differences between MSDs of b estimates for the beta ability distribution and MSDs of b estimates for the other ability distributions. In moving from the normal to the beta ability distribution, the MSD for the JML estimates increased more than did the MSD for the MB or the MML estimates. The b estimates for the normal and the truncated normal ability distributions tend to have similar MSDs. These trends are depicted in Figures 4, 5, and 6.

Accuracy of the c Parameter Estimation

For the normal and the truncated normal ability distributions, the median correlations were highest for the MB estimates of the c parameters (see Table 12). The median correlations for the MML estimates of the c parameters were either lower than or equal to those for the JML estimates. This same pattern occurs for the MSDs in Table 12. However, for the normal and truncated normal distributions the MSDs suggested smaller accuracy differences between the MB and JML estimates than were indicated by the correlations. The MML estimates were consistently less accurate than the JML or the MB estimates of the c parameters according to the MSDs. Examination of scatterplots in Figures 7, 8, and 9 also indicates that, for the normal and truncated normal distributions, the c parameters were best estimated by the MB procedure. Points in the MML scatterplot are more

TABLE 12

Accuracy Indices for c Parameter Estimates: 20 Items and 250 Examinees.

Estimation Procedure	Correlation Percentiles			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
MML	.41	.50	.58	0.019	0.002	0.017
MB	.65	.71	.78	0.012	0.004	0.008
JML	.62	.66	.71	0.013	0.004	0.009
Truncated θ						
MML	.49	.59	.73	0.020	0.005	0.015
MB	.67	.72	.84	0.013	0.004	0.009
JML	.54	.59	.71	0.015	0.004	0.011
Beta θ						
MML	.45	.48	.54	0.022	0.004	0.018
MB	.67	.70	.76	0.020	0.009	0.011
JML	.67	.72	.79	0.012	0.003	0.009

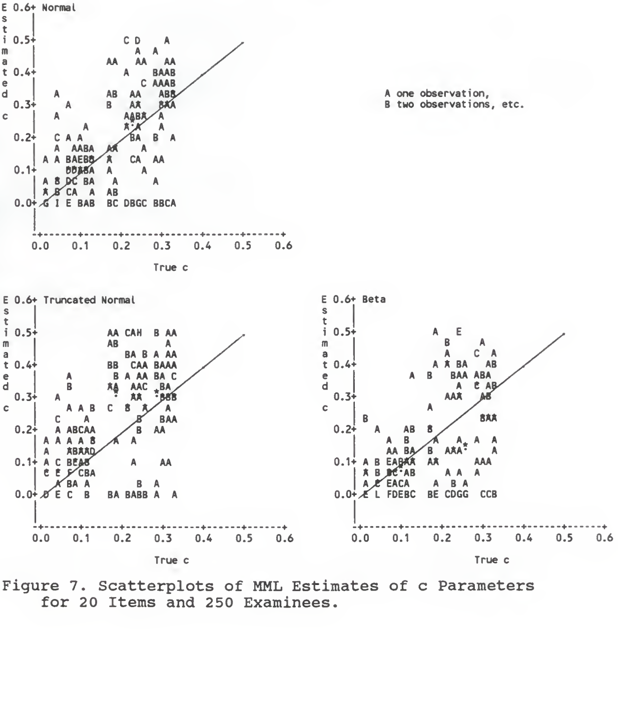


Figure 7. Scatterplots of MML Estimates of c Parameters for 20 Items and 250 Examinees.

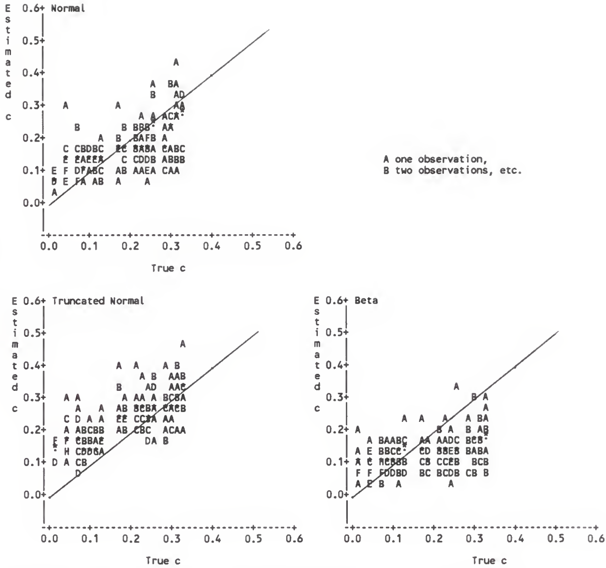


Figure 8. Scatterplots of MB Estimates of c Parameters
for 20 Items and 250 Examinees.

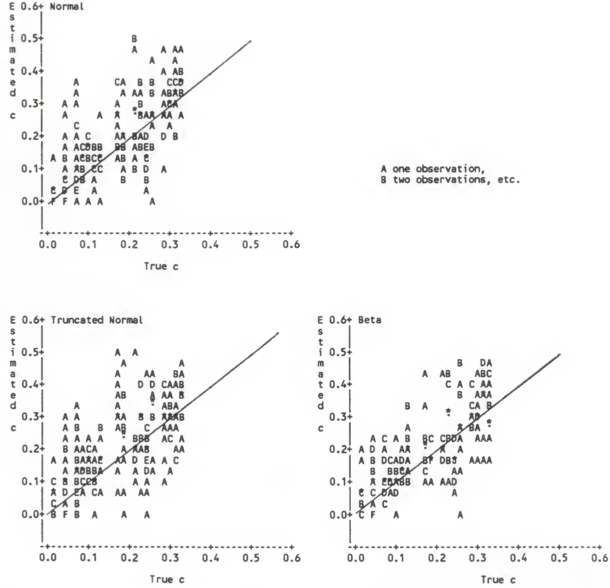


Figure 9. Scatterplots of JML Estimates of c Parameters for 20 Items and 250 Examinees.

scattered away from the agreement line than are the points in the scatterplots for the JML and MB estimates.

Except when the JML procedure was used, the magnitude of bias was smallest with the normal ability. As reported in Table 12, the JML estimates of the c parameters were equally biased for the normal and the truncated normal ability distributions. Both values were larger than the value of bias produced with the beta ability distribution.

With the beta distribution, the JML estimates were more accurate than the MB estimates. The latter were more accurate than the MML estimates. The primary differences between the JML and MB estimates was in the bias at the upper end of the c parameter scale (see Figure 8). There the MB procedure appeared to be more biased.

With the MML and the MB estimation procedures, the MSDs were lowest for the normal ability distribution. The effect of ability distribution on estimation of c parameters was negligible except for one condition. In moving from the normal to the beta ability distribution, the MSD for the MB estimates of the c parameters increased by approximately twofold (see Table 12).

Accuracy of the θ Parameter Estimation

The median correlations were similar for the ML-MB and ML-MML estimates of θ; both were higher than the median correlation for the JML estimates of θ parameters (see

Table 13). The MSDs for the θ parameter estimates followed the same pattern. The MSDs for the JML estimates were three times as large as the MSDs for either the ML-MML or the ML-MB estimates of the θ parameters. The reason for the large MSDs for the JML estimates can be seen by inspecting Figure 12: There is a substantial negative bias for low true θs. This bias is not evident in the scatterplots for the ML-MB or the ML-MML estimates (see Figures 10 and 11).

MSDs in Table 13 were similar across ability distributions except for two conditions: The MSD for the ML-MML and ML-MB procedures increased in moving from the normal to the truncated normal ability distributions. Similarly, the median correlations increased within each of the two procedures in moving from the truncated normal to the normal ability distribution. These trends are depicted in Figures 10, 11, and 12. The reason for increased accuracy in moving from truncated normal to normal ability distributions can be seen by comparing corresponding scatterplots in Figures 10 and 11. In Figure 10, the scatterplot for truncated normal ability indicates more negative bias than for normal ability distribution.

Plots of MSDs at several true ability levels are presented in Figures 13, 14, and 15 for normal, truncated normal, and beta ability distributions, respectively. There were clear differences among the three estimation procedures

TABLE 13

Accuracy Indices for θ Parameter Estimates: 20 Items and 250 Examinees.

Estimation Procedure	Correlation Percentiles			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
ML-MML	.88	.89	.89	0.740	0.232	0.508
ML-MB	.90	.90	.90	0.748	0.170	0.578
JML	.75	.76	.77	3.814	0.895	2.919
Truncated θ						
ML-MML	.83	.84	.85	0.968	0.271	0.697
ML-MB	.85	.86	.87	1.048	0.238	0.810
JML	.72	.74	.71	3.800	0.779	3.021
Beta θ						
ML-MML	.85	.86	.86	0.899	0.301	0.598
ML-MB	.86	.88	.89	0.867	0.207	0.660
JML	.77	.79	.79	3.730	1.009	2.721

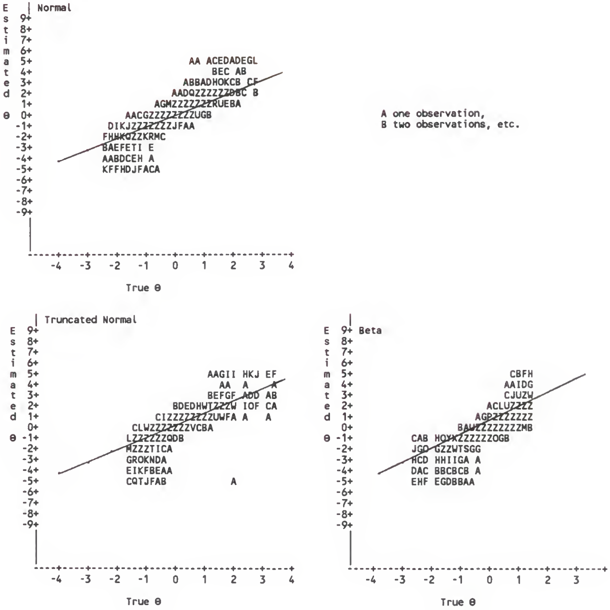


Figure 10. Scatterplots of ML-MML estimates of θ parameters for 20 items and 250 examinees.

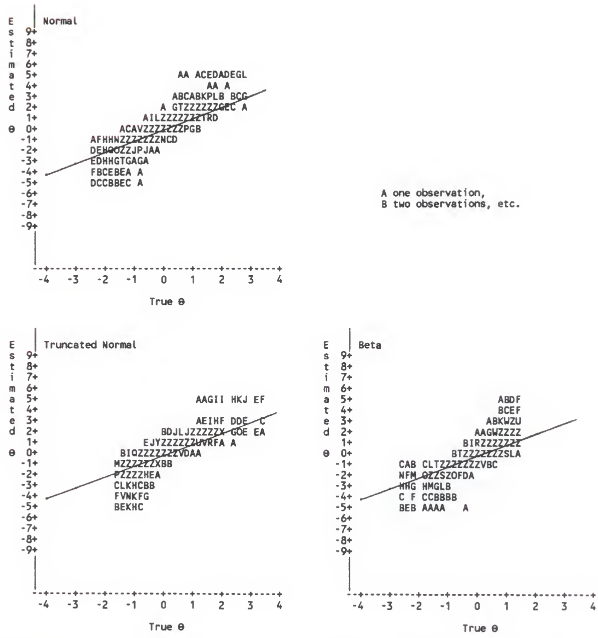


Figure 11. Scatterplots of ML-MB estimates of θ parameters for 20 items and 250 examinees.

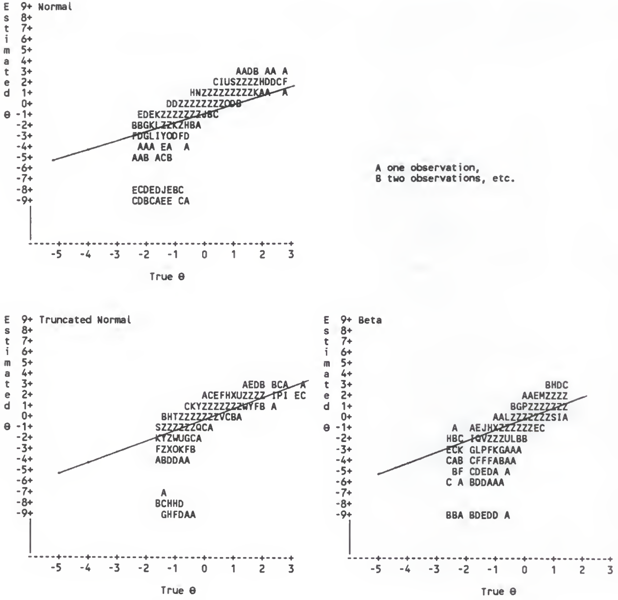


Figure 12. Scatterplots of JML estimates of θ parameters for 20 items and 250 examinees.

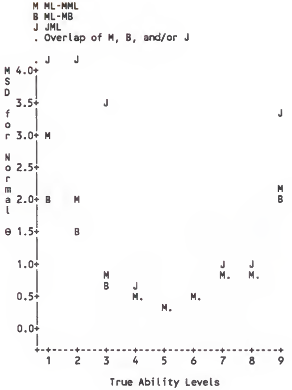


Figure 13. Plot of MSDs Versus Ability Levels: 20 Item, 250 Examinees, and Normal Ability Distribution.

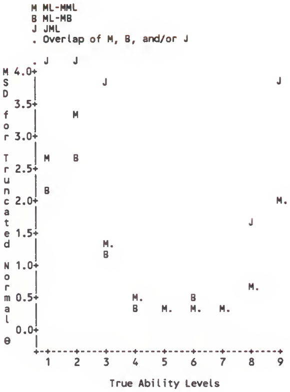


Figure 14. Plot of MSDs Versus Ability Levels: 20 Item, 250 Examinees, and Truncated Ability Distribution.

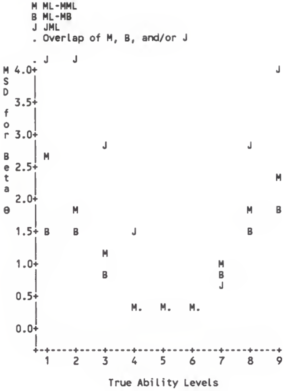


Figure 15. Plot of MSDs Versus Ability Levels: 20 Item, 250 Examinees, and Beta Ability Distribution.

within each of the three ability distributions. The differences were consistently at the upper and the lower levels of the ability distribution. At the upper and the lower levels, the ML-MB ability estimates were the most accurate and the JML ability estimates were the least accurate.

Plots of the MSDs at several true ability levels are presented in Figures 16, 17, and 18 for the ML-MML, ML-MB, and JML estimates respectively. The Plots indicate relatively little effect of ability distribution on accuracy of ability estimation except at the lower end of the ability distribution. There the truncated normal distribution tended to result in less accurate estimation.

Short Tests and Large Sample Sizes

For 20 items and 1000 examinees, the estimation accuracy indices are reported in Tables 14, 15, 16, and 17.

Accuracy of the a Parameter Estimation

An examination of the correlations in Table 14 indicated that within each ability distribution, the MB procedure had the highest median correlations, followed by the MML, and finally the JML procedure. Similarly, the MSDs for the MB procedure were smaller than for the other two procedures. Squared bias and variance were also smaller. With the normal distribution the MML procedure had the smallest bias.

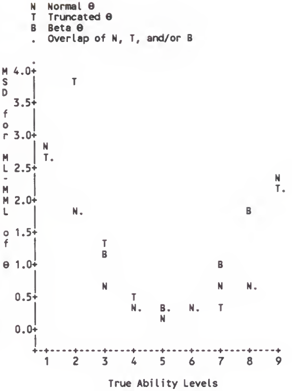


Figure 16. Plot of MSDs Versus Ability Levels: 20 Item, 250 Examinees, and ML-MML Estimation Procedure.

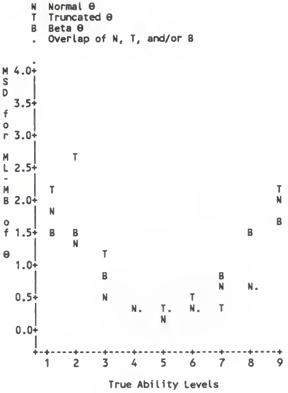


Figure 17. Plot of MSDs Versus Ability Levels: 20 Item, 250 Examinees, and MB-MML Estimation Procedure.

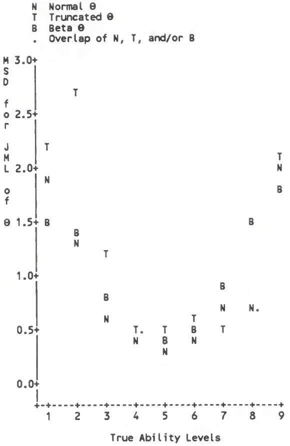


Figure 18. Plot of MSDs Versus Ability Levels: 20 Item, 250 Examinees, and JML Estimation Procedure.

TABLE 14

Accuracy Indices for a Parameter Estimates: 20 Items and 1000 Examinees.

Estimation Procedure	Correlation Percentiles			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
MML	.89	.89	.90	0.161	0.031	0.130
MB	.89	.90	.90	0.063	0.021	0.042
JML	.45	.50	.56	0.371	0.153	0.218
Truncated θ						
MML	.86	.87	.87	0.156	0.040	0.116
MB	.86	.88	.88	0.104	0.038	0.066
JML	.47	.52	.57	0.449	0.201	0.248
Beta θ						
MML	.86	.87	.88	0.336	0.128	0.208
MB	.87	.88	.88	0.320	0.154	0.166
JML	.26	.29	.37	0.508	0.218	0.290

In agreement with the results for the correlations, the MSDs for the MML estimates were smaller than those for the JML estimates of the a parameters within each ability distribution. The difference in MSDs for the MML and the MB is largely due to differences in variance: The bias of the two procedures was approximately the same within ability distributions. The scatterplots for different ability distributions and estimation procedures are presented in the same order as in the section for small sample size and short test. The reason for the smaller MSDs for the MB procedure than for the other procedures can be seen by comparing corresponding scatterplots in Figure 19, 20, and 21. In Figure 20, for the MB estimation procedure, the points were not scattered as widely as the points in the scatterplots for the MML (Figure 19) and JML (Figure 21) procedures.

For the MML and the MB the median correlations were very similar across the three ability distributions. For the JML procedure, the lowest median correlation was observed when the beta ability distribution was used. The correlations for the normal and the truncated normal ability distributions were similar; both were larger than for the beta ability distribution. For the three procedures, the MSDs decreased in moving from beta to the truncated normal, to the normal ability distribution. The decreases for the JML procedure were less appreciable than for the MB or the MML.

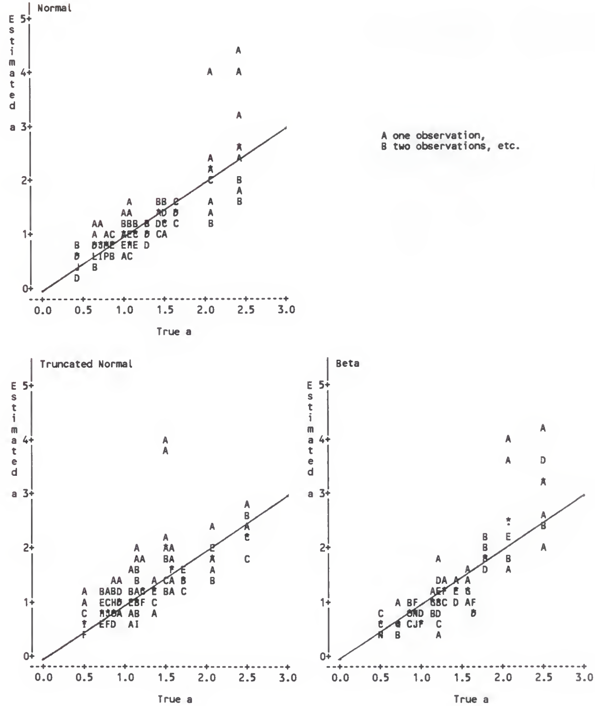


Figure 19. Scatterplots of MML Estimates of a Parameters for 20 Items and 1000 Examinees.

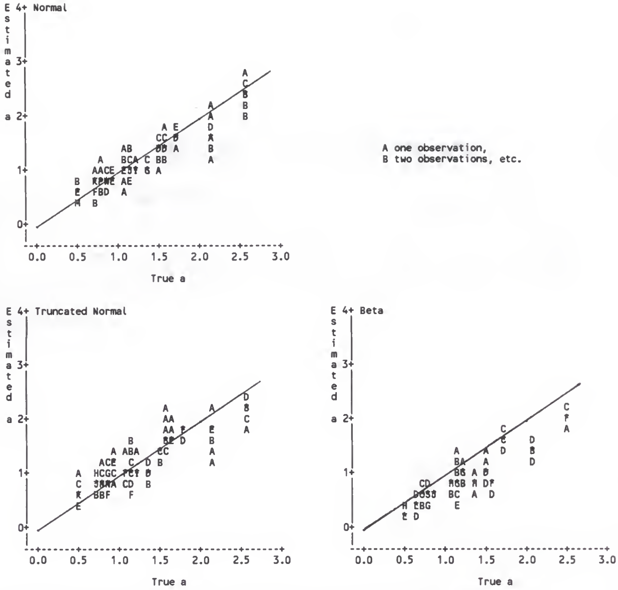


Figure 20. Scatterplots of MB Estimates of a Parameters for 20 Items and 1000 Examinees.

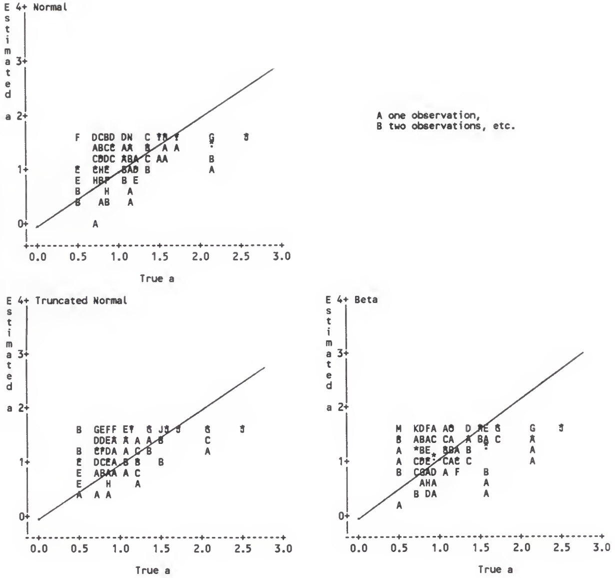


Figure 21. Scatterplots of JML Estimates of a Parameters for 20 Items and 1000 Examinees.

Accuracy of the b Parameter Estimation

Within the normal and the truncated normal ability distribution, the median correlations for the b parameter estimates were highest for the MB estimation procedure as indicated in Table 15. Within the beta ability distribution, the median correlations for the b parameter estimates were highest for the JML estimation procedure. For each ability distribution, the MSDs were lowest for the MB estimates and highest for the JML estimates. The MSDs for the MB estimates of the b parameters were similar to those for the MML estimates. The MSDs of the JML were 9 times as high as that of the MML or the MB estimates of the b parameters, within each ability distribution. Both the bias and the variance components of the MSD are larger for the JML estimates than for the MB or MML estimates. The reason for the increased MSDs of the JML estimates of the b parameters can be seen by comparing corresponding scatterplots in Figures 22, 23, and 24.

For each of the three estimation procedures, there were no appreciable differences in the median correlations across the three ability distributions. In the MSDs, there were small differences among ability distributions. For the MML and MB estimation procedures, the normal ability distribution had the lowest MSD. For the JML estimation

TABLE 15

Accuracy Indices for b Parameter Estimates: 20 Items and 1000 Examinees.

Estimation Procedure	Correlation Percentiles			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
MML	.86	.88	.93	0.143	0.031	0.112
MB	.92	.94	.95	0.120	0.050	0.070
JML	.97	.97	.98	5.340	2.631	0.709
Truncated θ						
MML	.85	.86	.86	0.432	0.206	0.226
MB	.97	.98	.99	0.430	0.205	0.225
JML	.96	.96	.97	4.338	2.147	2.191
Beta θ						
MML	.64	.70	.79	0.249	0.076	0.173
MB	.81	.85	.88	0.231	0.154	0.077
JML	.97	.97	.97	5.475	2.694	2.781

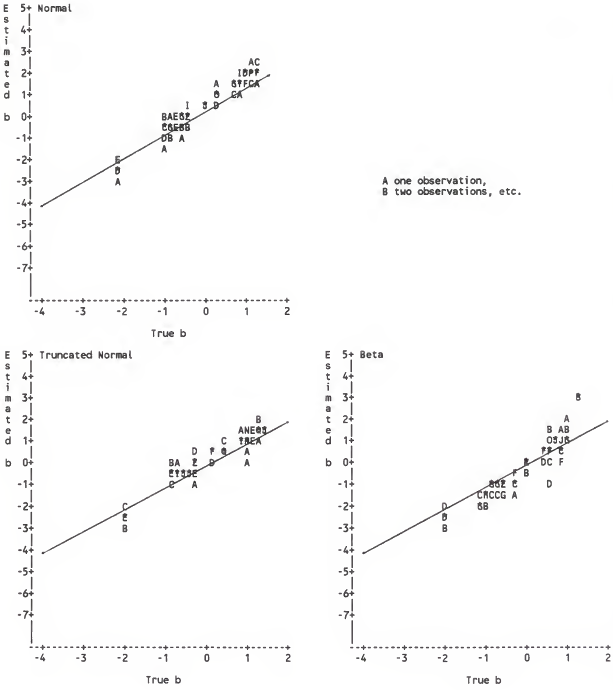


Figure 22. Scatterplots of MML Estimates of b Parameters for 20 Items and 1000 Examinees.

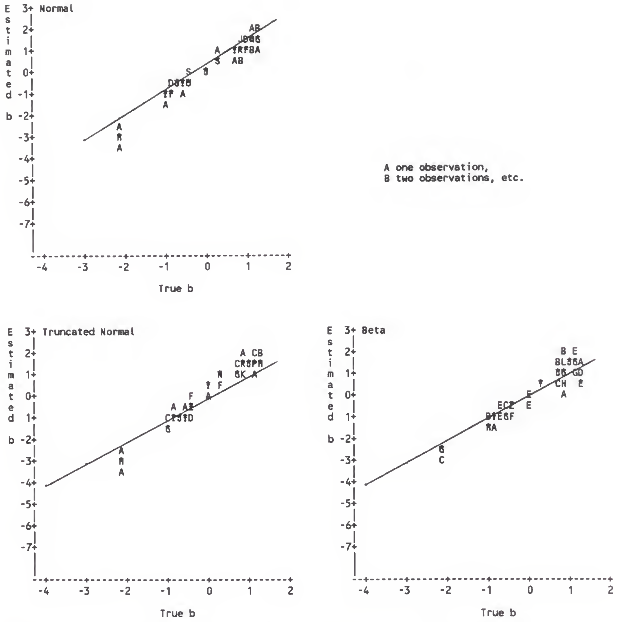


Figure 23. Scatterplots of MB Estimates of b Parameters for 20 Items and 1000 Examinees.

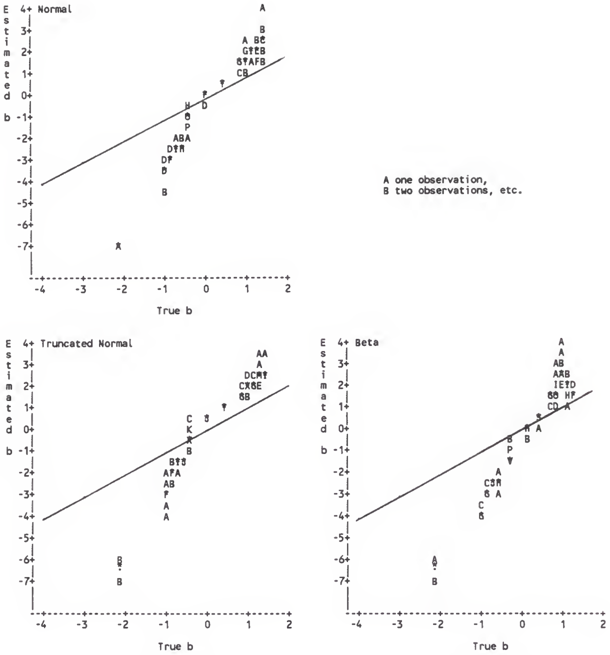


Figure 24. Scatterplots of JML Estimates of b Parameters
for 20 Items and 1000 Examinees.

procedure, the truncated normal ability distribution had the lowest MSD.

Accuracy of the c Parameter Estimation

For each ability distribution the median correlations were similar for the MML and MB estimates; both were higher than for the JML estimates of the c parameters (see Table 16). For the normal and the truncated normal ability distributions, the MSDs were lowest for MB estimates of the c parameters. For the beta ability distribution, the MSDs were lowest for the JML estimates of the c parameters and largest for the MML estimates. Within the normal ability distribution, the MSD for the MML estimates was lower than for the JML estimates. Within the truncated normal ability distribution, the MSD for the MML estimates was equal to that for the JML estimates; both were higher than for the MB estimates. Examination of scatterplots in Figures 25, 26, and 27 also indicates that for the normal and truncated normal ability distributions, the c parameters were best estimated by the MB procedure. Points in the MB scatterplot for the MB estimates are more evenly scattered above and below the agreement line than are the points in the scatterplots for the MML or JML estimates.

Except when the JML procedure was used, the magnitude of bias was smallest with the normal ability distribution, as reported in Table 16. The JML estimates produced similar

TABLE 16

Accuracy Indices for c Parameter Estimates: 20 Items and 1000 Examinees.

Estimation Procedure	Correlation Percentiles			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
MML	.76	.83	.92	0.010	0.001	0.009
MB	.81	.88	.94	0.006	0.002	0.004
JML	.54	.60	.61	0.019	0.008	0.011
Truncated θ						
MML	.95	.97	.98	0.020	0.008	0.012
MB	.97	.97	.98	0.010	0.004	0.006
JML	.47	.60	.70	0.020	0.009	0.011
Beta θ						
MML	.68	.78	.83	0.020	0.006	0.014
MB	.73	.80	.85	0.017	0.008	0.009
JML	.63	.67	.70	0.014	0.006	0.008

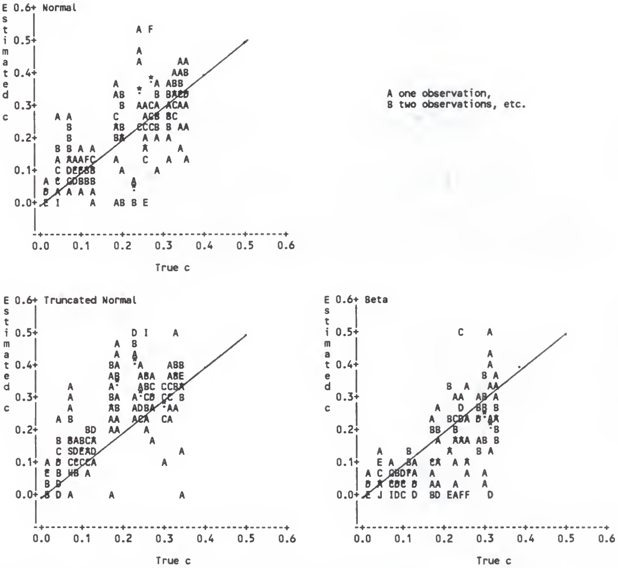


Figure 25. Scatterplots of MML Estimates of c Parameters for 20 Items and 1000 Examinees.

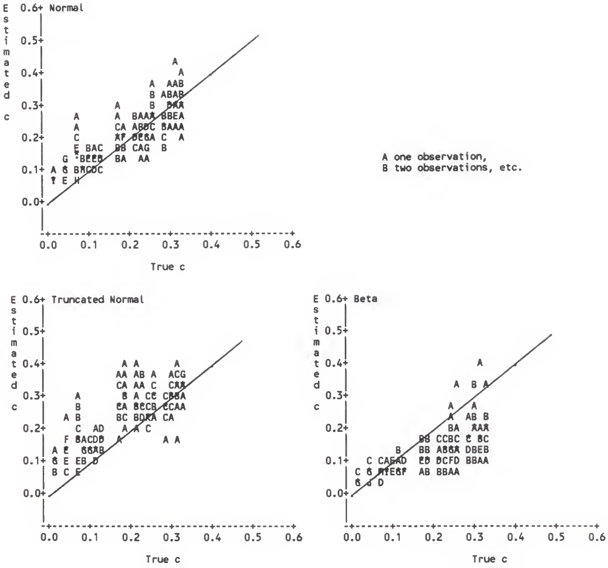


Figure 26. Scatterplots of MB Estimates of c Parameters for 20 Items and 1000 Examinees.

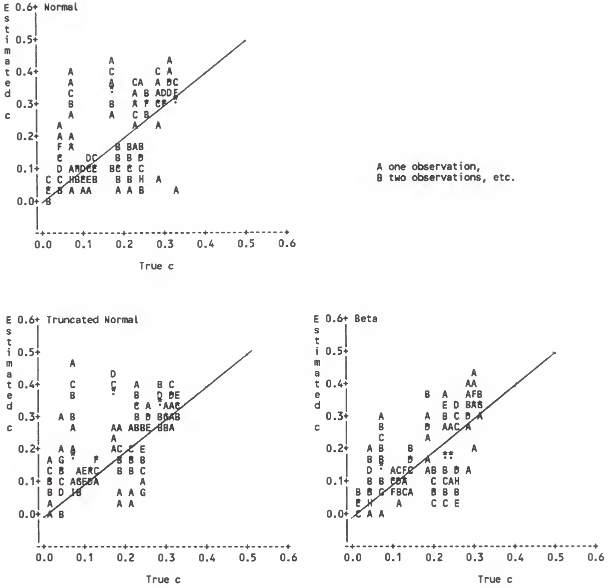


Figure 27. Scatterplots of JML Estimates of c Parameters for 20 Items and 1000 Examinees.

levels of bias for both normal and truncated normal ability distributions. Both values were larger than the value of bias produced with the beta ability distribution.

For the MML and the MB estimation procedures, the MSDs were lowest for the normal ability distribution. The effect of ability distribution on estimation of c parameter was more appreciable for the MML and MB than for the JML estimates. The MSD for the MB estimates increased in moving from the normal to the truncated normal and from truncated normal to beta ability distribution. The MSD for the MML estimates increased by twofold in moving from normal to truncated or beta ability distribution.

Accuracy of the θ Parameter Estimation

The median correlations were similar for the ML-MB and ML-MML estimates of θ; both were higher than the median correlation for the JML estimates of θ parameters except when the ability distribution was beta (see Table 17). Within the beta ability distribution, median correlations were similar for ML-MB and JML estimates; both were higher than for ML-MML estimates of θ parameters. The MSDs for the θ parameter estimates followed the same pattern except when the truncated normal ability distribution was used. With this distribution, the MSDs for ML-MML and ML-MB estimates were similar; both were higher than the MSD for the JML estimation procedure. With the normal ability distribution,

TABLE 17

Accuracy Indices for θ Parameter Estimates: 20 Items and 1000 Examinees.

Estimation Procedure	Correlation Percentiles			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
ML-MML	.87	.90	.92	1.048	0.296	0.752
ML-MB	.88	.91	.94	0.641	0.148	0.493
JML	.80	.81	.83	1.156	0.244	0.912
Truncated θ						
ML-MML	.90	.95	.97	1.237	0.305	0.932
ML-MB	.96	.97	.98	1.221	0.230	0.991
JML	.77	.80	.80	1.066	0.191	0.875
Beta θ						
ML-MML	.50	.58	.72	1.061	0.298	0.763
ML-MB	.76	.81	.84	1.035	0.288	0.747
JML	.85	.86	.86	1.029	0.253	0.776

the MSDs for the JML estimates were larger than for the ML-MML or the ML-MB estimates of θ parameters. The reason for the large MSDs for the JML estimates can be seen by inspecting Figure 30: There is negative bias for low true θs. This bias is not evident in the scatterplots for the ML-MB or the ML-MML estimates (see Figures 28 and 29). The reason for smaller MSDs for the ML-MB than for the ML-MML or the JML estimates of the normal ability distribution, can be seen by comparing corresponding scatterplots in Figures 28, 29, and 39. In Figure 29, the points in the scatterplot of the ML-MB estimates of the normal ability distribution are more evenly scattered above and below the agreement line than in the other two figures.

Within each estimation procedure, the median correlations were similar for the normal and the truncated normal ability distributions. For the ML-MB and the ML-MML, the correlations were smaller for the beta ability distribution. For the JML, the median correlation was largest with the beta distribution. For ML-MML and JML estimation procedures, the MSDs were also similar across ability distributions. The MSD for the ML-MB was smallest for the normal ability distribution. This small MSD is depicted in Figure 29. Points in the scatterplot for the normal ability distribution are more evenly scattered than for the other two distributions.

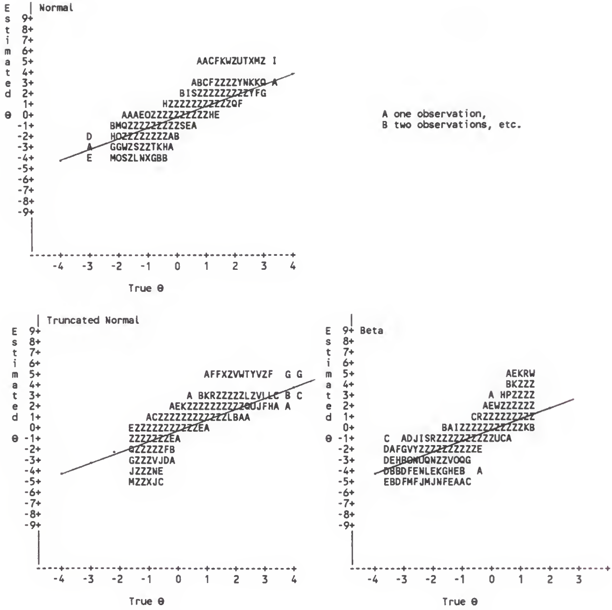


Figure 28. Scatterplots of ML-MML estimates of θ parameters for 20 items and 1000 examinees.

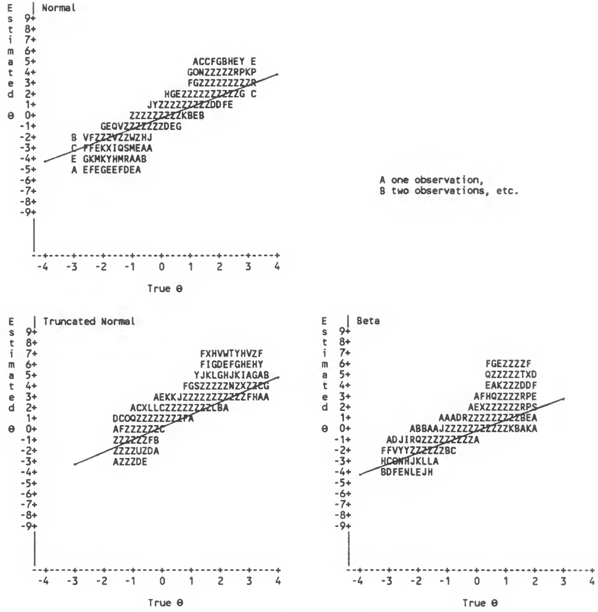


Figure 29. Scatterplots of ML-MB estimates of θ parameters for 20 items and 1000 examinees.

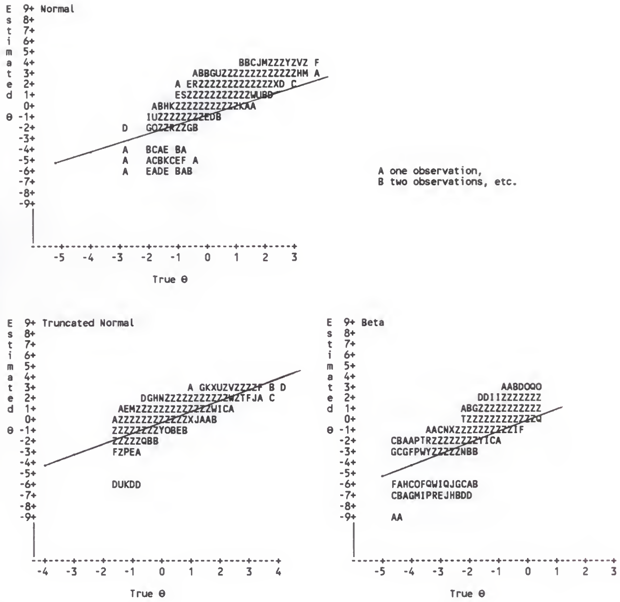


Figure 30. Scatterplots of JML estimates of θ parameters for 20 items and 1000 examinees.

Plots of MSDs at several true ability levels are presented in Figures 31, 32, and 33 for normal, truncated normal, and beta ability distributions, respectively. There were clear differences among the three estimation procedures with each of the three ability distributions. The differences were consistently at the upper and the lower levels of the ability distribution. The JML ability estimates were most accurate at the upper levels but least accurate at the lower levels of each of the three ability distributions. The accuracies of the ML-MB and the ML-MML were similar at the upper and the lower levels; both were most accurate at the upper levels but least accurate at the lower levels of θ. These accuracy differences at the upper and the lower levels of θ were more evident for the beta ability distribution (see Figure 33).

Plots of the MSDs at several true ability levels are presented in Figures 34, 35, and 36 for the ML-MML, ML-MB, and JML estimates respectively. The plots indicate relatively little effect of ability distribution on accuracy of ability estimation for ML-MML and JML estimates. The beta ability distribution resulted in less accurate estimation at the upper levels; the truncated ability indicated less accurate estimation at the lower levels of θ (see Figure 35). The JML was most accurate at the upper

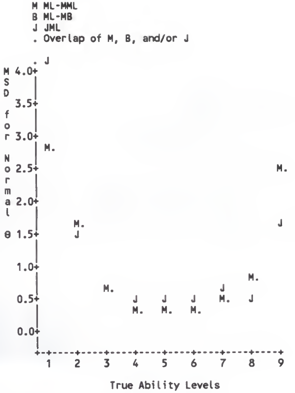


Figure 31. Plot of MSDs Versus Ability Levels: 20 Item, 1000 Examinees, and Normal Ability Distribution.

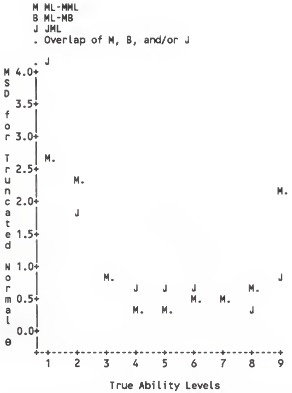


Figure 32. Plot of MSDs Versus Ability Levels: 20 Item, 1000 Examinees, and Truncated Ability Distribution.

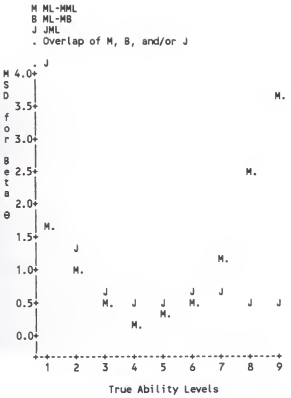


Figure 33. Plot of MSDs Versus Ability Levels: 20 Item, 1000 Examinees, and Beta Ability Distribution.

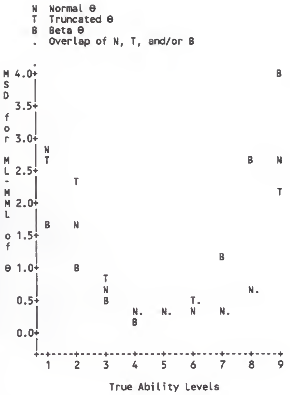


Figure 34. Plot of MSDs Versus Ability Levels: 20 Item, 1000 Examinees, and ML-MML Estimation Procedure.

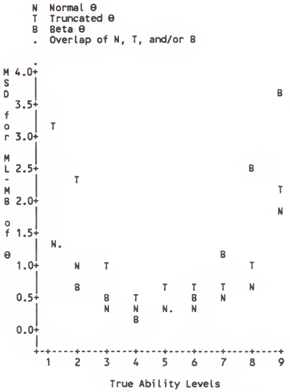


Figure 35. Plot of MSDs Versus Ability Levels: 20 Item, 1000 Examinees, and ML-MB Estimation Procedure.

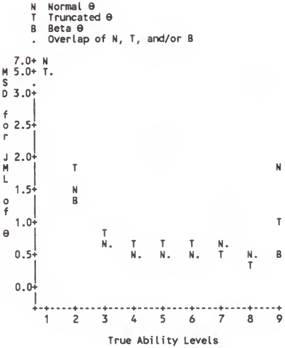


Figure 36. Plot of MSDs Versus Ability Levels: 20 Item, 1000 Examinees, and JML Estimation Procedure.

levels, whereas the ML-MML and the ML-MB were more accurate at the lower levels of the three ability distributions.

Short Tests and Large Sample Sizes

For 60 items and 250 examinees, the estimation accuracy indices are reported in Tables 18, 19, 20, and 21.

Accuracy of the a Parameter Estimation

An examination of the correlations in Table 18 indicated that within each ability distribution, the MB procedure had the highest median correlations, followed by the JML, and finally the MML procedure. The MSDs for the MB and JML procedure were similar and both were smaller than the MSDs for the MML estimates of the a parameters. The scatterplots for different ability distributions and estimation procedures are presented in the same order as in the section for small sample size and short test. The reason for the larger MSDs for the MML procedure can be seen by comparing Figures 37, 38, and 39. In Figure 37, the points were less evenly scattered above and below the agreement line than in the corresponding plots of Figure 38 and 39.

For the MML estimates of the a parameters, the median correlations were similar across the normal and the truncated normal ability distributions; both were lower than the median correlation for the beta ability distribution. For the MB estimates, the median correlations were similar

TABLE 18

Accuracy Indices for a Parameter Estimates: 60 Items and 250 Examinees.

Estimation Procedure	Correlation Percentile			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
MML	.42	.48	.54	0.816	0.193	0.623
MB	.78	.81	.83	0.125	0.045	0.080
JML	.57	.64	.69	0.163	0.038	0.201
Truncated θ						
MML	.39	.44	.53	1.167	0.351	0.816
MB	.69	.73	.79	0.149	0.050	0.099
JML	.67	.69	.73	0.150	0.030	0.120
Beta θ						
MML	.58	.64	.69	0.276	0.059	0.217
MB	.64	.68	.72	0.213	0.096	0.117
JML	.55	.62	.66	0.225	0.069	0.156

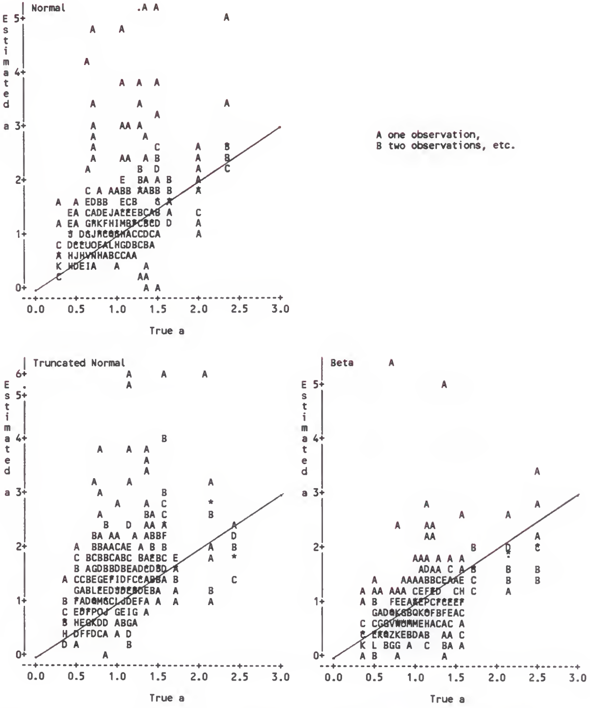


Figure 37. Scatterplots of MML Estimates of a Parameters for 60 Items and 250 Examinees.

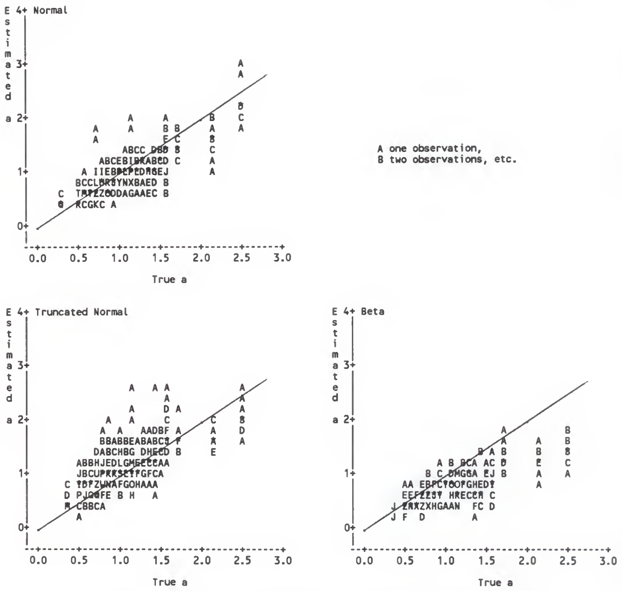


Figure 38. Scatterplots of MB Estimates of a Parameters for 60 Items and 250 Examinees.

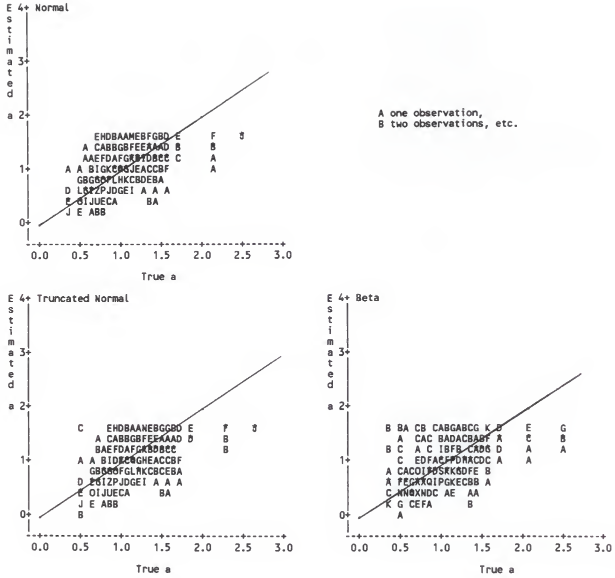


Figure 39. Scatterplots of JML Estimates of a Parameters for 60 Items and 250 Examinees.

for the truncated normal and the beta ability distributions; both were lower than the median correlation for the normal ability distribution. For the JML estimates, the median correlations were similar for the normal and the beta ability distributions; both were lower than that for the truncated normal ability distribution. These patterns also occurred in the MSDs. These differences in MSDs are depicted in Figures 37, 38, and 39.

Accuracy of the b Parameter Estimation

The median correlations for the b parameter estimates were highest for the MB procedure as indicated in Table 19. The median correlation was larger for the JML estimates than for the MML estimates except when the abilities had a beta distribution. Then the two median correlations were similar in magnitude. This same pattern occurs for the MSDs in Table 19. Both the bias and the variance components of MSD are smaller for the MB estimates than for the JML or MML estimates. Superiority of MB estimates, is depicted in Figures 40, 41, and 42. Points in scatterplots of Figure 41 for the MB estimates are more evenly scattered around the agreement line than in the scatterplots of Figures 40 and 42. For each of the estimation procedures, there were appreciable differences between MSDs of the b estimates for the normal ability distribution and MSDs of b estimates for the other ability distributions. In moving from the normal

TABLE 19

Accuracy Indices for b Parameter Estimates: 60 Items and 250 Examinees.

Estimation Procedure	Correlation Percentile			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
MML	.87	.90	.92	0.395	0.056	0.339
MB	.96	.96	.97	0.197	0.072	0.125
JML	.91	.93	.94	0.351	0.093	0.258
Truncated θ						
MML	.85	.88	.90	0.600	0.188	0.412
MB	.93	.94	.96	0.354	0.134	0.220
JML	.90	.92	.94	0.498	0.146	0.352
Beta θ						
MML	.80	.86	.90	1.180	0.280	0.900
MB	.95	.96	.97	0.333	0.202	0.282
JML	.87	.93	.94	1.174	0.383	0.791

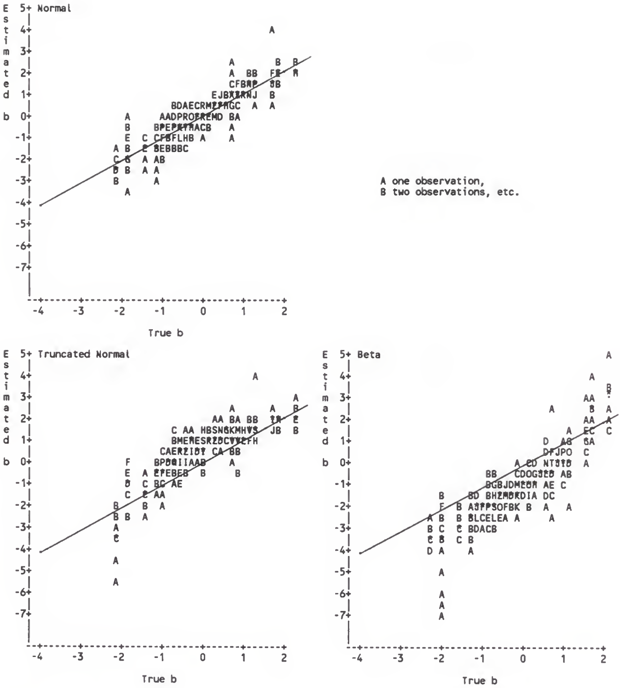


Figure 40. Scatterplots of MML Estimates of b Parameters for 60 Items and 250 Examinees.

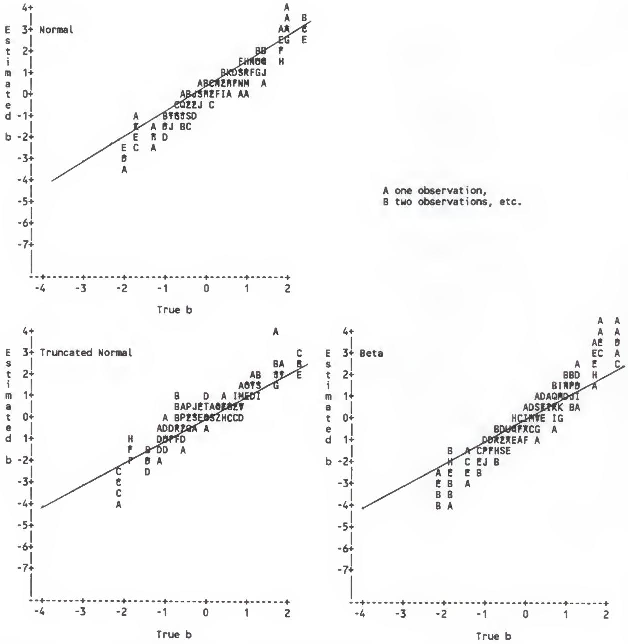


Figure 41. Scatterplots of MB Estimates of b Parameters for 60 Items and 250 Examinees.

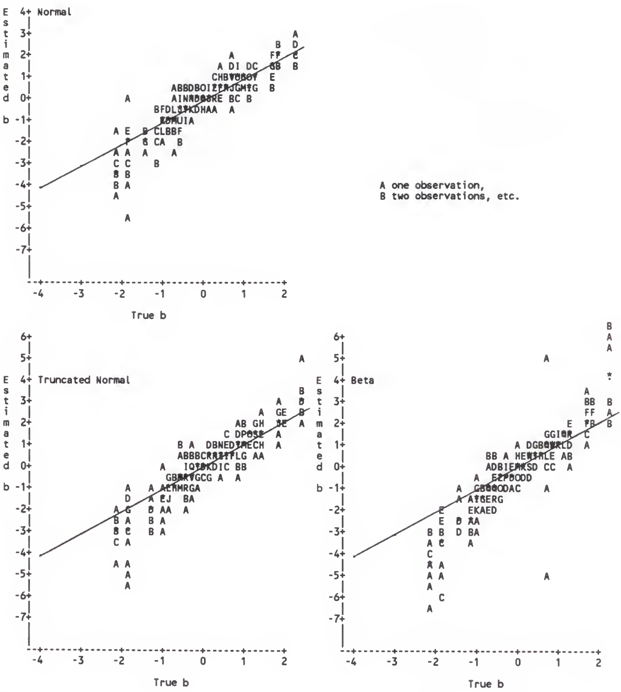


Figure 42. Scatterplots of JML Estimates of b Parameters for 60 Items and 250 Examinees.

to the truncated normal or beta ability distribution, the MSDs for the JML estimates increased more than did the MSDs for the MB or the Table 19 MML estimates. These trends are depicted in Figures 40, 41, and 42. The plots reveal that JML had the poorest estimates particularly at the lower end of difficulty scale. The JML estimates of the b parameters were also poor at the upper end, when the ability distribution was beta.

Accuracy of the c Parameter Estimation

For the three ability distributions, the median correlations were highest for the MB estimates of the c parameters (see Table 20). The median correlations for the MML estimates of the c parameters were either lower than or similar to those for the JML estimates. The MSDs for the MML were higher than those for the MB or the JML estimates. Examinations of scatterplots in Figures 43, 44, and 45 also indicates that the c parameters were best estimated by the MB procedure.

Except when the JML procedure was used, the magnitude of bias was smallest with the normal ability distribution. As reported in Table 20, the JML estimates of the c parameters were equally biased for the normal and the truncated normal ability distributions. Both values were smaller than the value of bias produced with the beta ability distribution.

TABLE 20

Accuracy Indices for c Parameter Estimates: 60 Items and 250 Examinees.

Estimation Procedure	Correlation Percentile			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
MML	.45	.52	.63	0.024	0.004	0.020
MB	.64	.68	.72	0.009	0.004	0.005
JML	.51	.56	.63	0.015	0.003	0.012
Truncated θ						
MML	.46	.57	.62	0.036	0.011	0.025
MB	.61	.64	.68	0.018	0.008	0.010
JML	.47	.58	.58	0.016	0.003	0.013
Beta θ						
MML	.33	.42	.48	0.029	0.005	0.024
MB	.60	.60	.67	0.016	0.007	0.009
JML	.52	.55	.57	0.016	0.004	0.012

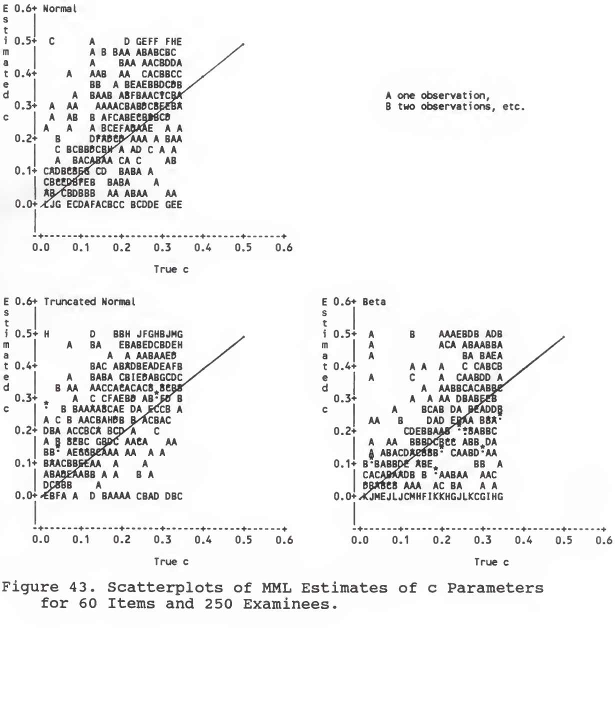


Figure 43. Scatterplots of MML Estimates of c Parameters for 60 Items and 250 Examinees.

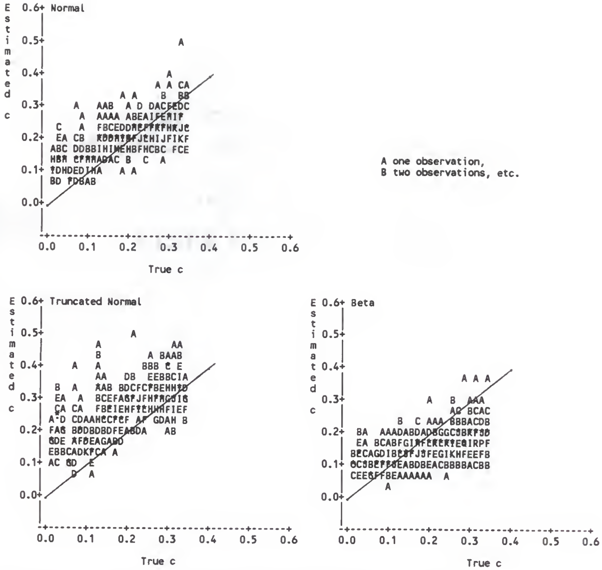


Figure 44. Scatterplots of MB Estimates of c Parameters for 60 Items and 250 Examinees.

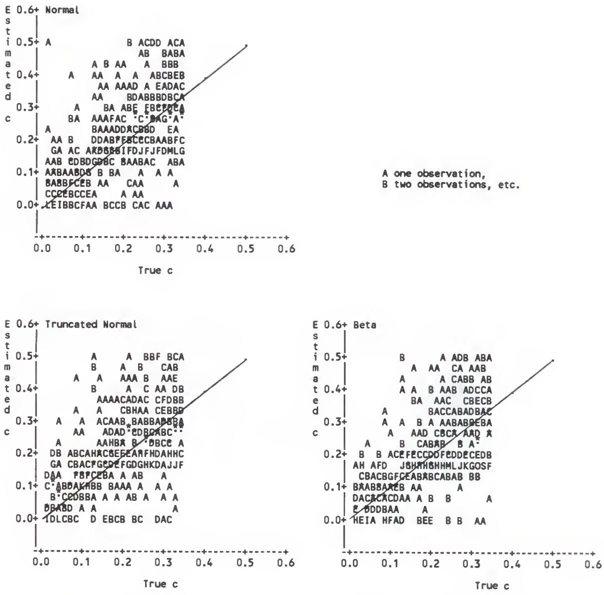


Figure 45. Scatterplots of JML Estimates of c Parameters for 60 Items and 250 Examinees.

The effect of ability distribution on estimation of c parameters was negligible except when the MB procedure was used. The MSD for the MB estimates of the normal ability distribution was smaller than the MSDs for the truncated normal or beta ability distributions. The reason for smaller MSD can be seen by comparing corresponding scatterplots in Figure 44.

Accuracy of the θ Parameter Estimation

For the normal and the beta ability distributions, the median correlations were highest for the ML-MB estimates of θ. With the truncated normal distribution the JML procedure had the highest median correlation (see Table 21). Accuracy, as indicated by MSDs, followed the same pattern. The MSDs for the ML-MML estimates were smaller than those for the JML estimates with the normal and beta ability distributions. For the truncated normal ability distribution, the MSD decreased by about sixfold in moving from the ML-MML estimates to the JML estimates; by twofold in moving from the ML-MML to ML-MB estimates. The reason for this decrease can be seen by inspecting corresponding scatterplots in Figures 46, 47, and 48. In Figure 46, there is a substantial negative bias for low and high values of the truncated normal ability distribution. This bias is not evident in the scatterplots for the ML-MB estimates (see

TABLE 21

Accuracy Indices for θ Parameter Estimates: 60 Items and 250 Examinees.

Estimation Procedure	Correlation Percentile			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
ML-MML	.85	.91	.92	0.541	0.106	0.435
ML-MB	.94	.95	.95	0.327	0.082	0.245
JML	.89	.91	.93	0.577	0.131	0.446
Truncated θ						
ML-MML	.50	.69	.89	1.323	0.282	1.041
ML-MB	.89	.91	.93	0.468	0.108	0.360
JML	.94	.96	.96	0.281	0.069	0.212
Beta θ						
ML-MML	.93	.94	.94	0.339	0.087	0.252
ML-MB	.94	.95	.95	0.405	0.032	0.373
JML	.93	.93	.94	0.610	0.199	0.411

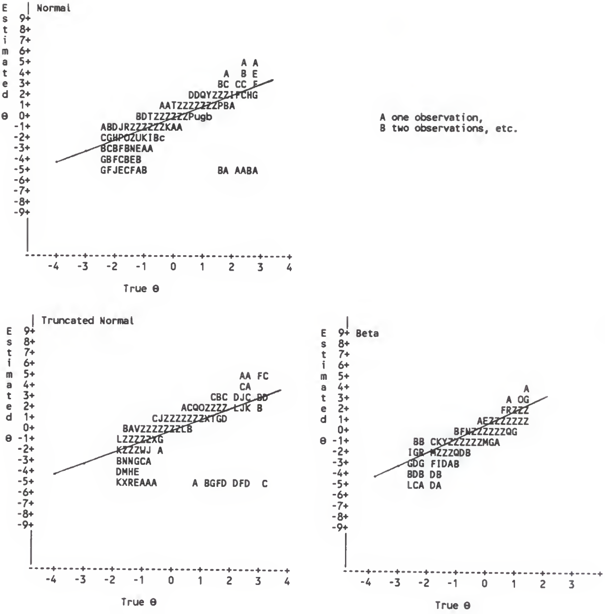


Figure 46. Scatterplots of ML-MML Estimates of θ Parameters for 60 Items and 250 Examinees.

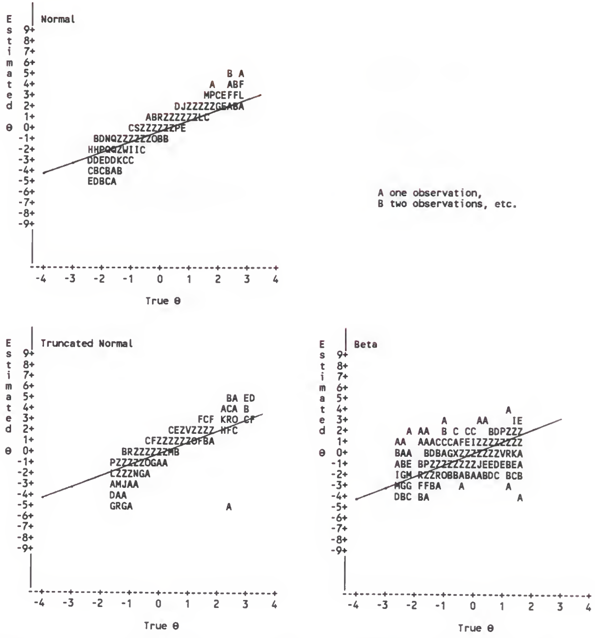


Figure 47. Scatterplots of ML-MB Estimates of θ Parameters for 60 Items and 250 Examinees.

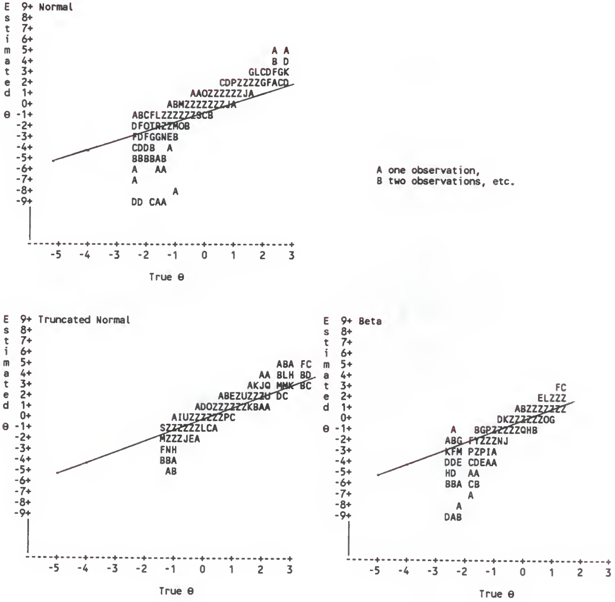


Figure 48. Scatterplots of JML Estimates of θ Parameters for 60 Items and 250 Examinees.

Figures 47). In Figure 48, bias was only in evidence for low values of the ability distributions.

MSDs for the ML-MB were similar across ability distributions. The MSDs for the ML-MML increased in moving from the beta to the normal ability distributions and from the normal to the truncated normal ability distributions. The MSDs for the JML increased in moving from the truncated normal to the normal and from the normal to the beta ability distribution. These trends are depicted in Figures 46 and 48. The reason for the increase in MSD for the ML-MML estimates, in moving from the beta to the truncated normal or normal ability distributions, can be seen by comparing corresponding scatterplots in Figures 46. In Figure 46, the scatterplot for the truncated normal ability indicates more negative bias than do the scatterplots for the beta or the normal ability distribution.

Plots of MSDs at several true ability levels are presented in Figures 49, 50, and 51 for normal, truncated normal, and beta ability distributions, respectively. There were clear differences among the three estimation procedures within the normal and the truncated normal ability distributions. These differences were consistently at the upper and the lower levels of the ability distribution. At the upper and the lower levels, the ML-MB and the JML

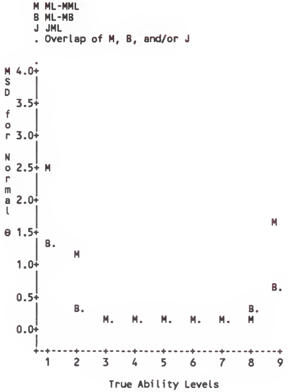


Figure 49. Plot of MSDs Versus Ability Levels: 60 Item, 250 Examinees, and Normal Ability Distribution.

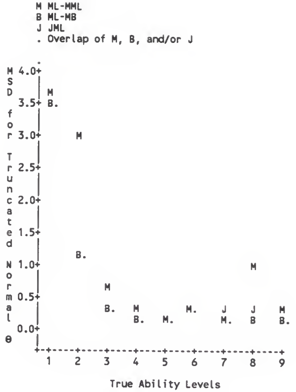


Figure 50. Plot of MSDs Versus Ability Levels: 60 Item, 250 Examinees, and Truncated Ability Distribution.

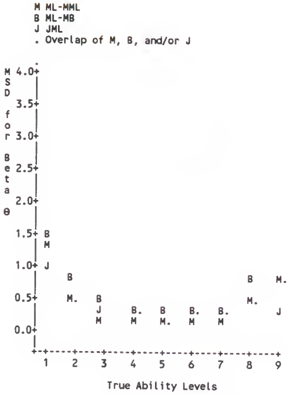


Figure 51. Plot of MSDs Versus Ability Levels: 60 Item, 250 Examinees, and Beta Ability Distribution.

ability estimates were more accurate than the ML-MML estimates of θ.

Plots of the MSDs at several true ability levels are presented in Figures 52, 53, and 54 for the ML-MML, ML-MB, and JML estimates respectively. The Plots indicate clear differences at the upper and/or the lower levels of the ability distributions. At the lower levels, the ML-MB estimates with the truncated normal ability distribution were less accurate than with the beta ability distribution. At the lower and the upper levels, the ML-MML estimates with the truncated normal ability distribution were less accurate than with the beta ability distribution. At the lower levels, the JML estimates with the truncated normal ability distribution were more accurate than with the normal or the beta ability distribution.

Large Tests and Large Sample Sizes

For 60 items and 1000 examinees, the estimation accuracy indices are reported in Tables 22, 23, 24, and 25.

Accuracy of the a Parameter Estimation

An examination of the correlations in Table 22 indicated that within each ability distribution, the three estimation procedures had similar median correlations. In disagreement with the results for the correlations, the MSDs for the JML estimates were smaller than those for the MML or the MB estimates of the a parameters, within the truncated normal

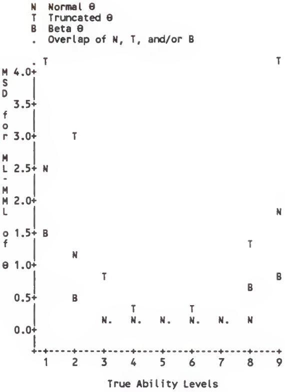


Figure 52. Plot of MSDs Versus Ability Levels: 60 Item, 250 Examinees, and ML-MML Estimation Procedure.

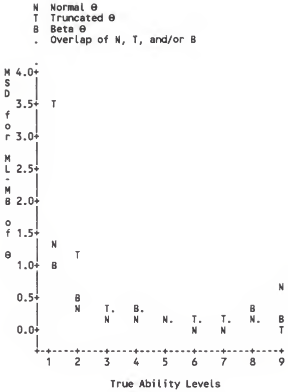


Figure 53. Plot of MSDs Versus Ability Levels: 60 Item, 250 Examinees, and MB-MML Estimation Procedure.

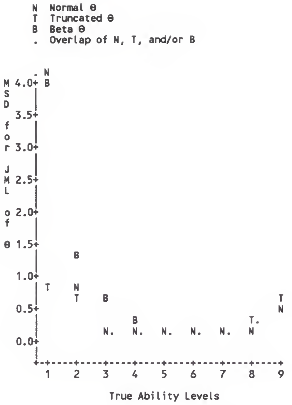


Figure 54. Plot of MSDs Versus Ability Levels: 60 Item, 250 Examinees, and JML Estimation Procedure.

TABLE 22

Accuracy Indices for a Parameter Estimates: 60 Items and 1000 Examinees.

Estimation Procedure	Correlation Percentile			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
MML	.86	.89	.90	0.095	0.090	0.086
MB	.91	.92	.94	0.045	0.015	0.030
JML	.85	.87	.88	0.091	0.034	0.057
Truncated θ						
MML	.82	.84	.87	0.174	0.064	0.110
MB	.82	.82	.89	0.184	0.078	0.106
JML	.86	.82	.88	0.073	0.026	0.047
Beta θ						
MML	.81	.84	.88	0.125	0.050	0.075
MB	.84	.86	.87	0.139	0.064	0.075
JML	.82	.84	.88	0.122	0.046	0.076

ability distribution. For the normal ability distribution, the MSD for the JML was smaller than for the MML but larger than for the MB estimates of the a parameters. The reason for smaller MSDs for the JML estimates can be seen by comparing corresponding scatterplots in Figures 55, 56, and 57. Points in scatterplots of Figure 57 are more evenly scattered around the agreement line than are the points in the corresponding scatterplots of Figure 55.

For the MML and the MB estimation procedure, the median correlations were highest when the ability distribution was normal. With the JML, the median correlations were similar across the ability distributions. This same pattern occurs for the MSDs. The reason for lower MSDs for the MML and MB estimates when the ability distribution was normal can be seen by comparing the corresponding scatterplots within Figure 55 and within Figure 56. Points in the scatterplots for the normal ability distribution are more evenly scattered around the agreement line than the points are in the non-normal ability distributions.

Accuracy of the b Parameter Estimation

For each of the three ability distributions, the median correlations were similar across the three estimation procedures as indicated in Table 23. In disagreement with the results for the correlation, the MSDs for the MB were smaller than those for the MML or the JML when the ability

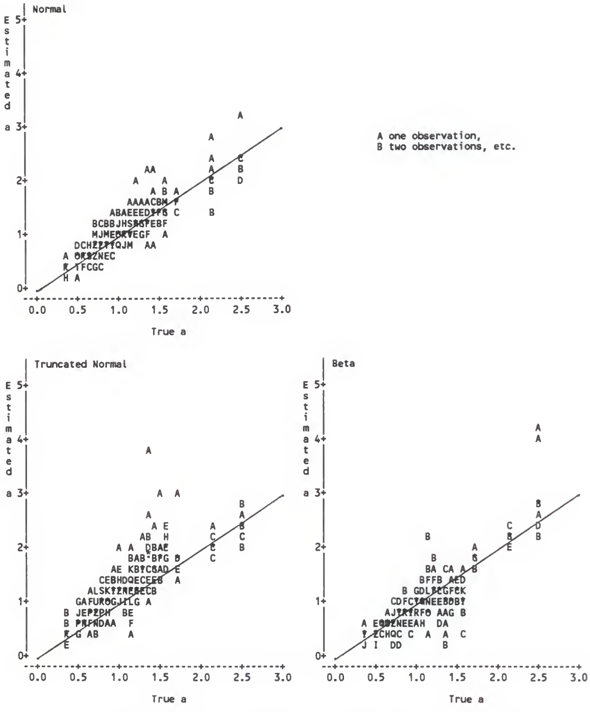


Figure 55. Scatterplots of MML Estimates of a Parameters for 60 Items and 1000 Examinees.

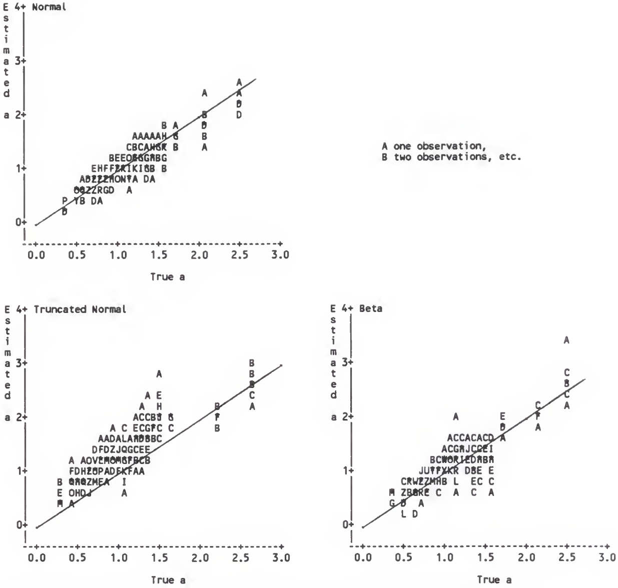


Figure 56. Scatterplots of MB Estimates of a Parameters for 60 Items and 1000 Examinees.

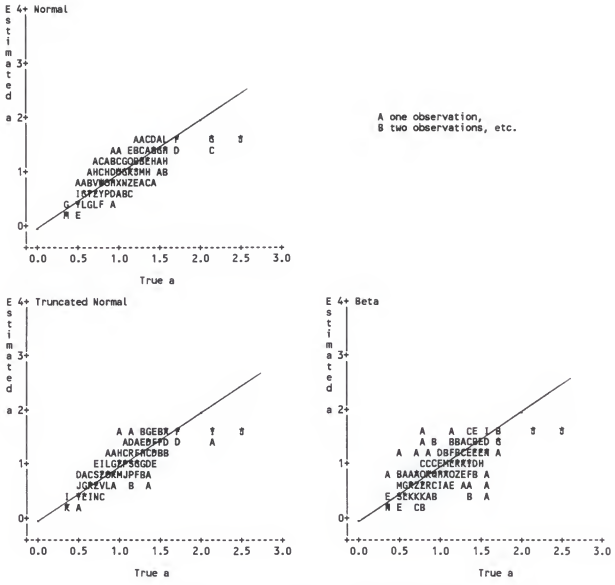


Figure 57. Scatterplots of JML Estimates of a Parameters for 60 Items and 1000 Examinees.

TABLE 23

Accuracy Indices for b Parameter Estimates: 60 Items and 1000 Examinees.

Estimation Procedure	Correlation Percentile			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
MML	.94	.95	.96	0.139	0.023	0.116
MB	.98	.98	.98	0.093	0.035	0.058
JML	.97	.97	.98	0.243	0.098	0.145
Truncated θ						
MML	.92	.94	.96	0.263	0.107	0.156
MB	.95	.96	.96	0.256	0.118	0.138
JML	.97	.98	.98	0.246	0.100	0.146
Beta θ						
MML	.92	.94	.95	0.291	0.095	0.196
MB	.96	.97	.97	0.205	0.087	0.118
JML	.96	.97	.97	0.441	0.179	0.262

distribution was normal or beta. The MSDs for the JML estimates were larger than those for MML estimates of the b parameters as were its components bias and variance. The reason for the smaller MSDs for the MB estimates can be seen by comparing corresponding scatterplots in Figures 58, 59, and 60. Points in the scatterplots of Figure 59, for the MB estimates are more evenly scattered around the agreement line than are the points in the corresponding plots of Figures 58 and 60.

For each of the three estimation procedures, the median correlations were similar (see Table 23). In disagreement with the correlation results, the MSDs for the b estimates increased in moving from the normal ability distribution to the truncated normal or the beta ability distribution. The reason for this increase can be seen by comparing corresponding scatterplots in Figures 58, 59, and 60. The points of the scatterplots for the normal ability distributions are more evenly scattered around the agreement line than for the other two distributions.

<u>Accuracy of the c Parameter Estimation</u>

For each ability distribution the median correlations were similar for the JML and MB estimates; both were higher than for the MML estimates (see Table 24). For the normal and the beta ability distributions, the MSDs were similar for the MB and the JML estimates; both were smaller

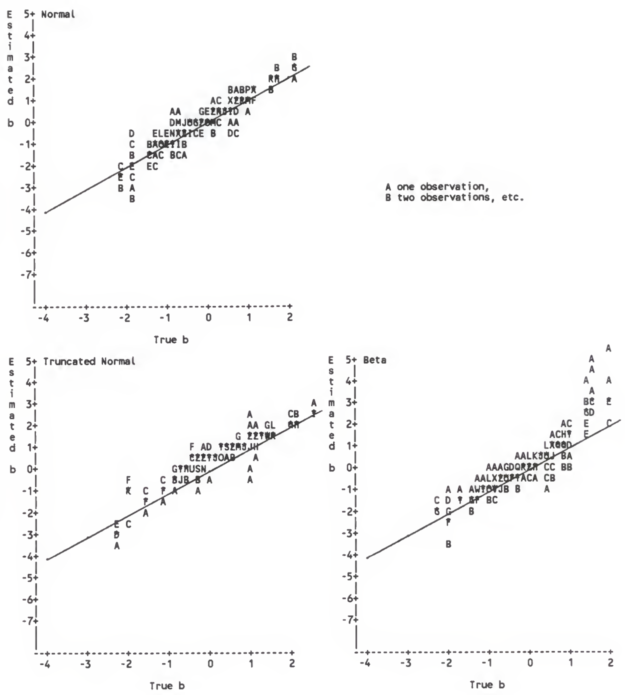


Figure 58. Scatterplots of MML Estimates of b Parameters for 60 Items and 1000 Examinees.

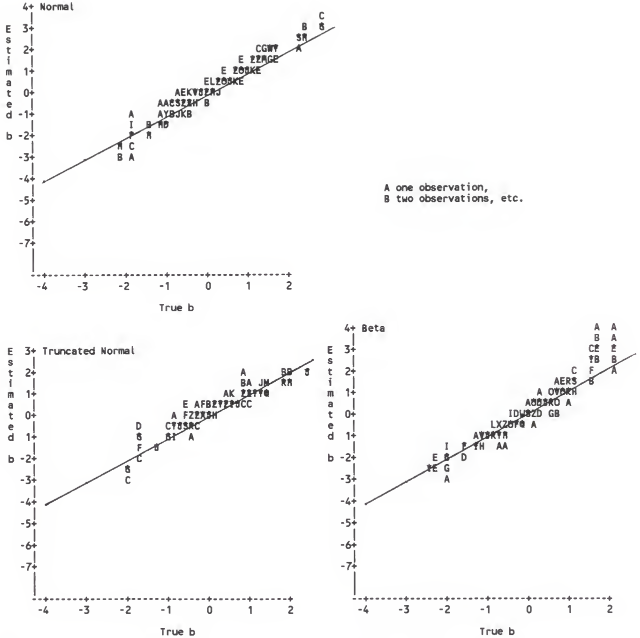


Figure 59. Scatterplots of MB Estimates of b Parameters for 60 Items and 1000 Examinees.

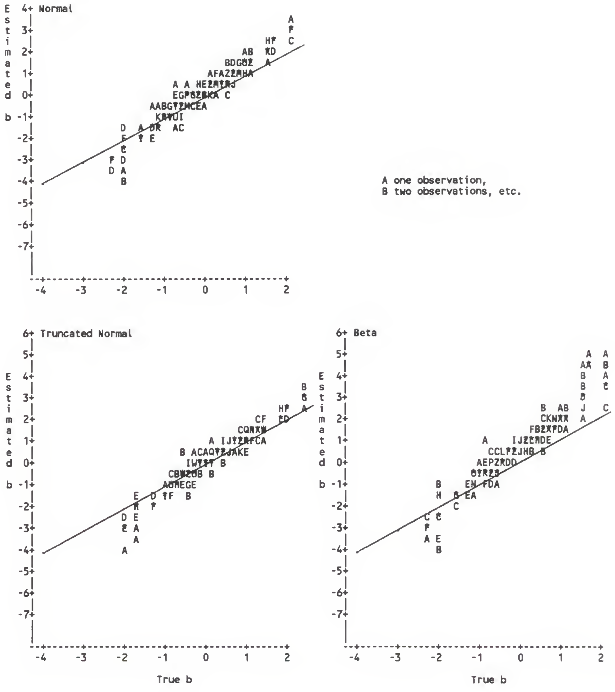


Figure 60. Scatterplots of JML Estimates of b Parameters for 60 Items and 1000 Examinees.

TABLE 24

Accuracy Indices for c Parameter Estimates: 60 Items and 1000 Examinees.

Estimation Procedure	Correlation Percentile			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
MML	.51	.61	.66	0.015	0.002	0.013
MB	.74	.76	.78	0.007	0.003	0.004
JML	.69	.71	.74	0.008	0.003	0.005
Truncated θ						
MML	.63	.67	.68	0.031	0.014	0.017
MB	.69	.71	.73	0.020	0.009	0.011
JML	.68	.69	.75	0.008	0.003	0.005
Beta θ						
MML	.59	.61	.67	0.019	0.006	0.013
MB	.73	.77	.80	0.012	0.005	0.007
JML	.75	.76	.79	0.007	0.002	0.004

than the MSDs for the MML estimates of the c parameters. For the truncated normal ability distribution the MSDs were similar for the MML and MB estimates of the c parameters; both were higher than the MSDs for the JML estimates.

Except when the JML procedure was used, the magnitude of bias was smallest with the normal ability. As reported in Table 24, the JML estimates produced similar values of bias with each of the three ability distributions.

With the MML and the MB estimation procedures, the median correlations were similar for the normal and beta ability distributions; both were higher than the median correlation for the truncated normal ability distribution. For the JML estimation procedure, the median correlations were similar for the three ability distributions. This same pattern occurs for the MSDs. These trends are depicted in Figures 64, 65, and 66.

Accuracy of the θ Parameter Estimation

For the three ability distributions, the median correlations were similar for the ML-MML, ML-MB, and JML estimates of θ as indicated in Table 25. The MSDs were similar for the ML-MML and ML-MB estimates under the normal and the beta ability distributions; both were smaller than the MSD for the JML estimation procedure (see Table 25). The reason for the increase of MSDs in moving from ML-MML or ML-MB to JML estimates can be seen by comparing

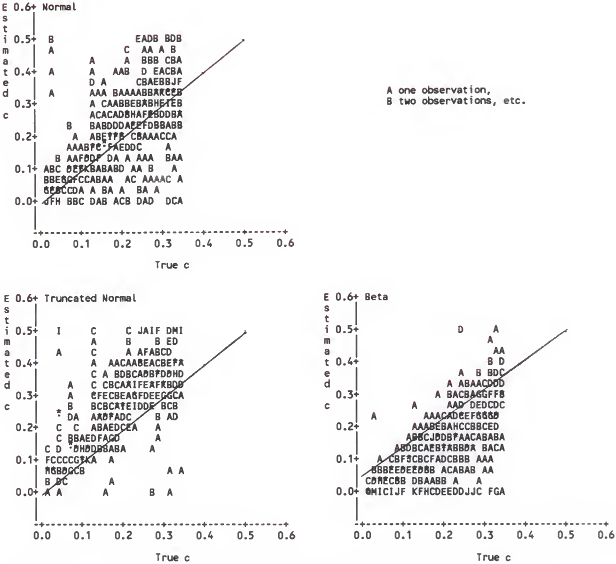


Figure 61. Scatterplots of MML Estimates of c Parameters for 60 Items and 1000 Examinees.

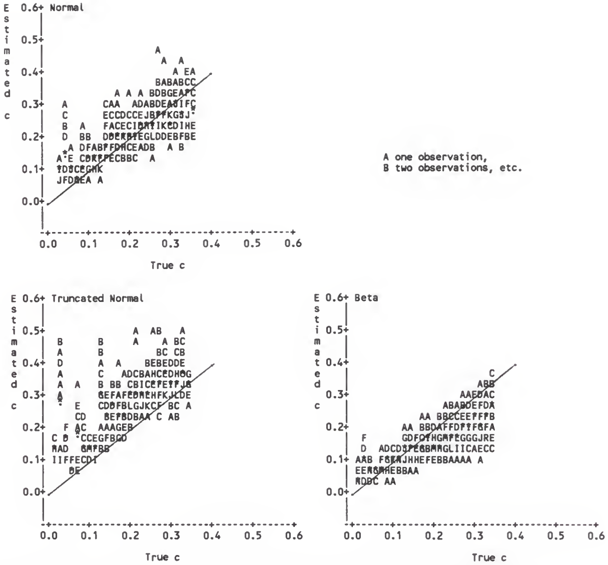
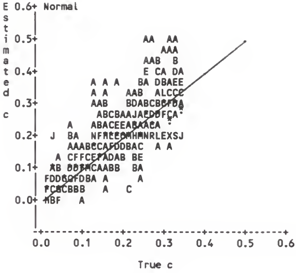
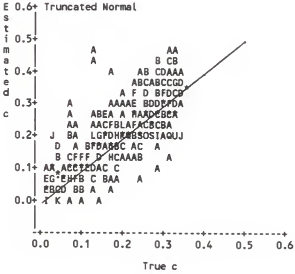


Figure 62. Scatterplots of MB Estimates of c Parameters for 60 Items and 1000 Examinees.



A one observation,
B two observations, etc.



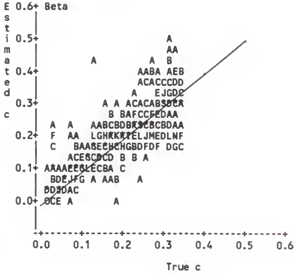


Figure 63. Scatterplots of JML Estimates of c Parameters for 60 Items and 1000 Examinees.

TABLE 25

Accuracy Indices for θ Parameter Estimates: 60 Items and 1000 Examinees.

Estimation Procedure	Correlation Percentile			MSD	Squared Bias	Variance
Normal θ	25th	50th	75th			
ML-MML	.94	.94	.95	0.229	0.045	0.184
ML-MB	.94	.94	.95	0.264	0.058	0.206
JML	.91	.93	.94	0.422	0.100	0.322
Truncated θ						
ML-MML	.87	.88	.89	0.570	0.138	0.432
ML-MB	.87	.88	.88	0.472	0.100	0.376
JML	.92	.94	.95	0.347	0.083	0.264
Beta θ						
ML-MML	.94	.94	.95	0.135	0.019	0.116
ML-MB	.94	.95	.95	0.148	0.026	0.122
JML	.91	.92	.92	0.618	0.183	0.435

corresponding scatterplots in Figures 64, 65, and 66. In Figure 66, the JML scatterplots for the normal and beta ability distributions indicate higher negative bias at the lower values of θ than the corresponding scatterplots indicate in Figures 64 and 65.

For the three estimation procedures, the median correlations were similar across the ability distributions (see Table 25). In disagreement with the results for the correlations, the MSDs for the ML-MML and ML-MB estimation increased in moving from the beta to normal and from normal to truncated normal ability distributions. For the JML estimation, the MSDs increased in moving from truncated normal to normal and from normal to beta ability distributions.

Plots of MSDs at several true ability levels are presented in Figures 67, 68, and 69 for normal, truncated normal, and beta ability distributions, respectively. There were clear differences among the three estimation procedures within each of the three ability distributions. The differences were consistently at the lower levels of the ability distribution. The ML-MB and the ML-MML had similar MSDs at the lower levels of the normal and truncated normal θ; both were lower than for the JML estimates of θ. For the beta ability distribution, the MSDs for the ML-MML and ML-MB were also similar; both were higher than for the JML

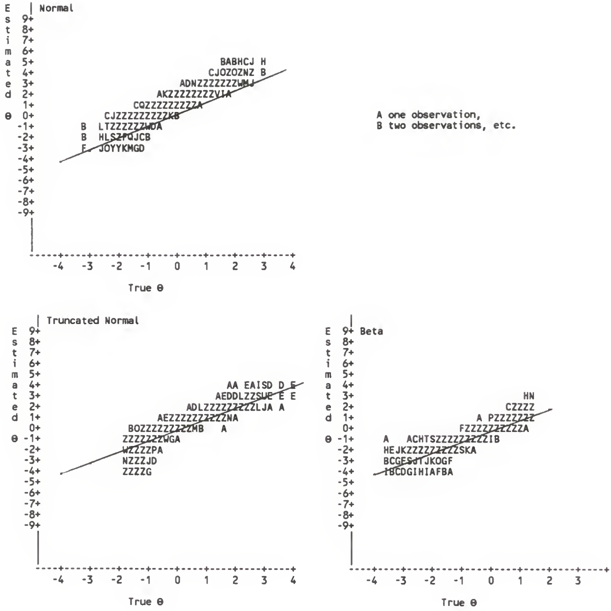


Figure 64. Scatterplots of ML-MML Estimates of θ parameters for 60 items and 1000 examinees.

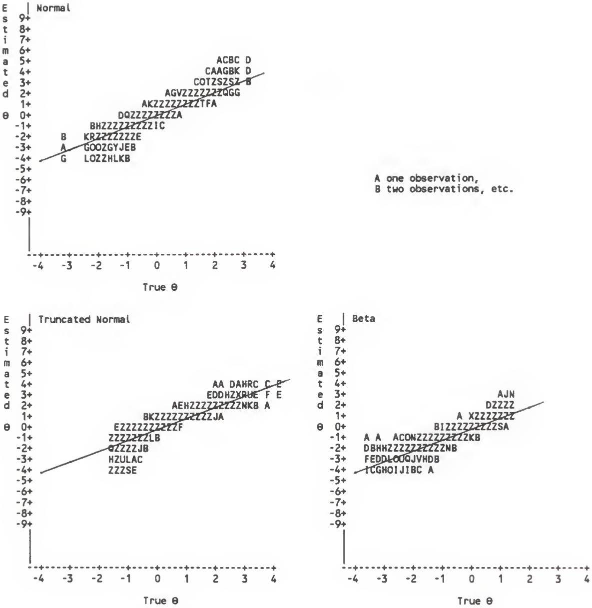


Figure 65. Scatterplots of ML-MB estimates of θ parameters for 60 items and 1000 examinees.

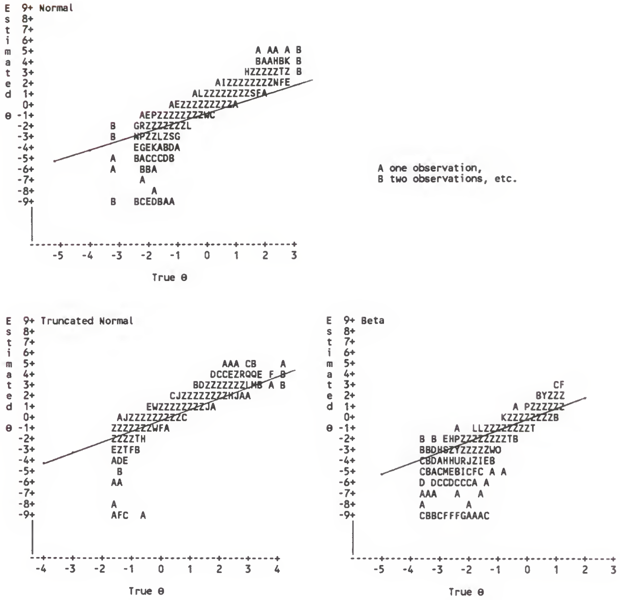


Figure 66. Scatterplots of JML estimates of θ parameters for 60 items and 1000 examinees.

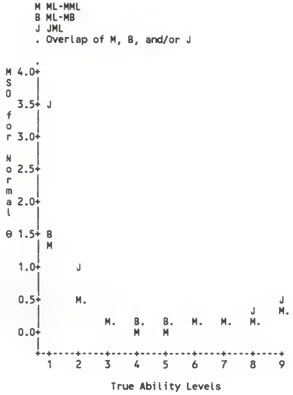


Figure 67. Plot of MSDs Versus Ability Levels: 60 Item, 1000 Examinees, and Normal Ability Distribution.

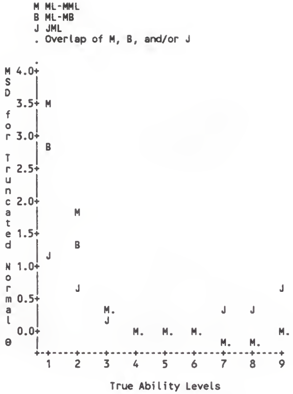


Figure 68. Plot of MSDs Versus Ability Levels: 60 Item, 1000 Examinees, and Truncated Ability Distribution.

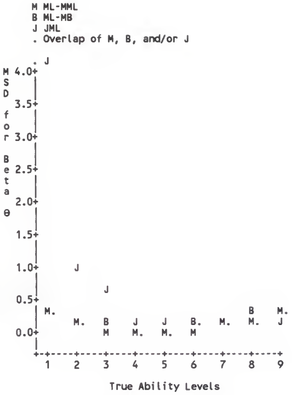


Figure 69. Plot of MSDs Versus Ability Levels: 60 Item, 1000 Examinees, and Beta Ability Distribution.

estimation. These accuracy differences at the lower levels of θ were more evident for the normal and truncated normal ability distribution (see Figures 67 and 68).

Plots of the MSDs at several true ability levels are presented in Figures 70, 71, and 72 for the ML-MML, ML-MB, and JML estimates respectively. The plots indicate effects of ability distribution on accuracy of ability estimation only at the lower levels of θ. For the ML-MML and the ML-MB estimation, the MSDs increased in moving from beta to normal and from normal to truncated normal ability distribution (see Figure 70 and 71). For the JML, the MSDs increased from truncated normal to normal and from normal to beta ability distribution (see Figure 72).

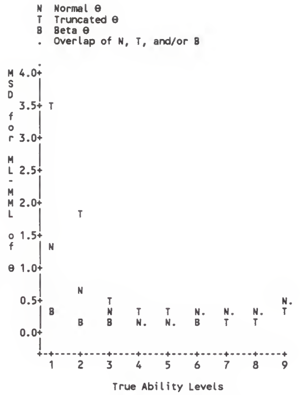


Figure 70. Plot of MSDs Versus Ability Levels: 60 Item, 1000 Examinees, and ML-MML Estimation Procedure.

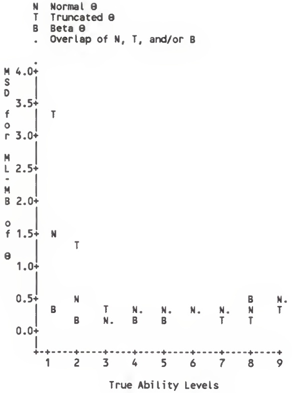


Figure 71. Plot of MSDs Versus Ability Levels: 20 Item, 1000 Examinees, and MB-MML Estimation Procedure.

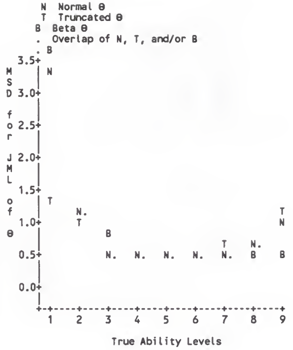


Figure 72. Plot of MSDs Versus Ability Levels: 60 Item, 1000 Examinees, and JML Estimation Procedure.

CHAPTER V
DISCUSSION AND CONCLUSION

In this study, the estimation accuracy for the three-parameter logistic model was investigated under two sample sizes, two test lengths, and three ability distributions. The JML of LOGIST and the MML and the MB of BILOG were used in calibrating each of 10 replications for the dichotomous data. Several criteria for estimation accuracy were used. Among these, MSD is the most important and bias is the second most important. The results of this investigation with respect to MSD and bias are summarized in the following sections. In the following, the term accuracy implies a comparison in terms of MSDs. Bias comparisons are reported only when MSDs were approximately the same but differences in bias occurred.

The results for the a parameters indicated following:

1. With the normal ability distribution, the MB estimates of the a parameters were more accurate (had lower MSDs) than the MML or the JML estimates except in one condition. With 60 items and 250 examinees the MB estimates were more biased than the JML estimates. The JML estimates were more accurate than the MML estimates, with two exceptions: (a) With 20 items and 1000 examinees the JML estimates were less accurate than the MML estimates, and

(b) with 20 items and 250 examinees the JML estimates were more biased than the MML estimates.

2. With the truncated normal ability distribution, the MB estimates of the a parameters were more accurate than the MML or the JML estimates with the following exceptions: (a) With 60 items and 250 examinees the JML estimates were as accurate as the MB estimates, and (b) with 60 items and 1000 examinees the MB estimates were as accurate as the MML estimates but less accurate than the JML estimates. The JML estimates were more accurate than the MML estimates, with two exceptions: (a) With 20 items and 1000 examinees the JML estimates were less accurate than the MML estimates, and (b) with 20 items and 250 examinees the JML estimates were as biased as the MML estimates.

3. With the beta ability distribution, the MB estimates of the a parameters were more accurate than the MML or the JML estimates, with several exceptions: (a) With 20 items and 250 examinees the MB estimates were more biased than the MML or the JML estimates, (b) with 20 items and 1000 examinees the MB estimates were more biased than the MML estimates, (c) with 60 items and 250 examinees the MB estimates were more biased than the MML or the JML estimates, and (d) With 60 items and 1000 examinees the MB estimates were less accurate than the JML and the MML estimates. The JML estimates were more accurate than the

MML with the following exception: (a) With 20 items and 250 examinees the JML estimates were more biased, (b) with 20 items and 1000 examinees the MML estimates were more accurate than the JML estimates, and (c) with 60 items and 1000 examinees the MML estimates were as accurate as the JML estimates.

4. The MB estimates of the a parameters were more accurate with the normal ability distribution than with the truncated normal or the beta ability distributions. The MB estimates were more accurate with the truncated normal ability than with the beta ability distribution except with 60 items and 1000 examinees.

5. The MML estimates of the a parameters were similar in accuracy across ability distributions, with several exceptions: (a) With 20 items and 250 examinees the MML estimates were less accurate with the normal ability distribution, (b) with 60 items and 250 examinees the MML estimates were more accurate with the beta ability distribution, (c) with 20 items and 1000 examinees the MML estimates were less accurate with the beta ability distribution, and (d) with 60 items and 1000 examinees the MML estimates were more accurate with the normal ability distribution.

6. The JML estimates of the a parameters were largely unaffected by the ability distribution. However, the JML

estimates tended to be more accurate with the normal ability when there were 20 items and 1000 examinees.

The results for the b parameters indicated following:

1. With the normal ability distribution, the MB estimates of the b parameters were more accurate than the MML or the JML estimates with the exception that the MML estimates tended to be less biased than the MB and the JML estimates. The JML estimates were less accurate than the MML estimates, with one exception. With 60 items and 250 examinees the JML estimates and the MML estimates were similar in accuracy.

2. With the truncated normal ability distribution, the MB estimates of the b parameters were more accurate than the MML or the JML estimates with 250 examinees. With 1000 examinees, the MB estimates were as accurate as the MML estimates. The JML estimates were less accurate than the other estimates when there were 20 items; however, the JML estimates were more accurate than the MML estimates with 60 items and 250 examinees, and as accurate as the MML estimates for 60 items and 1000 examinees.

3. With the beta ability distribution, the MB estimates of the b parameters were more accurate than the MML or the JML estimates, with two exceptions: (a) With 20 items and 250 examinees the MB estimates were as accurate as the MML estimates, and (b) with 20 items and 1000 examinees the MB

estimates were less accurate than the MML estimates. The MML estimates were more accurate than the JML estimates except for one case. With 60 items and 250 examinees the JML estimates were as accurate as the MML estimates.

4. The MB estimates of the b parameters were more accurate with the normal ability distributions than with the non-normal ability distributions, and, with one exception, more accurate with the beta ability distribution than with the truncated normal ability distribution. With 20 items and 250 examinees the MB estimates were more accurate with the truncated normal ability distribution than with the beta ability distribution.

5. The MML estimates of the b parameters were more accurate with the normal than with the non-normal ability distribution with the following exception: With 20 items and 250 examinees the MML estimates were similar in accuracy with the normal and truncated normal ability distributions. With one exception the MML estimates were more accurate with the truncated normal ability distribution than with the beta ability distribution. With 20 items and 1000 examinees the MML estimates were more accurate with the beta ability distribution than with the truncated normal ability distribution.

6. The JML estimates of the b parameters were more accurate with the normal than with the non-normal ability

distributions, with two exceptions: (a) With 20 items and 1000 examinees the JML estimates were similar in accuracy for the normal and the beta ability distributions, and (b) with 60 items and 1000 examinees the JML estimates were similar for the normal and the truncated normal ability distributions. The JML estimates were more accurate with the truncated normal ability distribution than with the beta ability distribution.

The results for the c parameters indicated following:

1. With the normal ability distribution the MB estimates of the c parameters were more accurate than the MML or the JML estimates, with two exceptions: (a) With 20 items and 250 examinees the MB estimates were more biased than the MML estimates and as accurate as the JML estimates, and (b) with 60 items and 1000 examinees the MB estimates were as accurate as the JML estimates. The JML estimates were more accurate than the MML estimates except in one condition. With 20 items and 1000 examinees the MML estimates were more accurate than the JML estimates.

2. With the truncated normal ability distribution the MB estimates of the c parameters were more accurate than the MML estimates. The JML estimates were more accurate than the MML estimates except in one condition. With 20 items and 1000 examinees the MML estimates were as accurate as the MML estimates.

3. With the beta ability distribution the JML estimates of the c parameters were more accurate than the MML or the MB estimates except in one condition. With 60 items and 250 examinees the JML estimates were as accurate as the JML estimates. The MML estimates were as accurate as the MB estimates with 20 items and more accurate than the MB estimates with 60 items.

4. The MB estimates were more accurate with the normal ability distribution than with the non-normal. The MB estimates with the truncated normal ability were as accurate or more accurate than with the beta ability distribution.

5. The MML estimates of the c parameters were more accurate with the normal ability distribution than with the non-normal ability distribution with one exception. With 20 items and 250 examinees the MML estimates were similar in accuracy with the three ability distributions. The MML estimates were more accurate with the beta ability distribution than with the truncated normal ability distribution when the number of items was 60. With 20 items the MML estimates were similar in accuracy with the beta and the truncated normal ability distribution.

6. The JML estimates of the c parameters were similar in accuracy across the three ability distributions.

The results for the θ parameters indicated following:

1. With the normal ability distribution the ML-MB estimates were more accurate than the ML-MML or the JML estimates except in two conditions: (a) With 20 items and 250 examinees the ML-MML and the ML-MB estimates were similar in accuracy, and (b) with 60 items and 1000 examinees the ML-MML and ML-MB estimates were similar in accuracy. The ML-MML estimates were more accurate than the JML estimates.

2. With the truncated normal ability distribution the ML-MB and the ML-MML estimates were similar in accuracy, with one exception: With 60 items and 250 examinees the ML-MB estimates were more accurate. Both were less accurate than the JML estimates except for one condition. With 20 items and 250 examinees the JML estimates were less accurate than the ML-MB and the ML-MML estimates.

3. With the beta ability distribution the ML-MB and the ML-MML estimates were similar in accuracy; both were more accurate than the JML estimates except in one condition. With 20 items and 1000 examinees the ML-MB and ML-MML estimates were similar in accuracy to the JML estimates.

4. The ML-MB estimates were most accurate with the normal ability distribution except in one condition. With 60 items and 1000 examinees the ML-MB estimates were most accurate with the beta ability distribution. With one

exception, the ML-MB estimates were more accurate with the beta ability distribution than with the truncated normal ability distribution except for one case. With 20 items and 1000 examinees the ML-MB estimates were more biased with the beta ability distribution than the truncated normal ability distribution.

5. The ML-MML estimates were most accurate for the normal ability distribution except in one condition. With 60 items and 1000 examinees the ML-MML estimates were most accurate with the beta ability distribution. The ML-MML estimates were more accurate with the beta ability distribution than with the truncated normal ability distribution except in one condition. With 20 items and 250 examinees the ML-MML estimates were similar in accuracy for the truncated normal and the beta ability distributions.

6. With 60 items the JML tended to be more accurate with the truncated normal ability distribution than with the normal or the beta ability distributions. With 20 items the JML estimates were more biased with the beta ability distribution than with the normal or the truncated normal ability distributions.

Thus accuracy of the estimation procedures depended upon the ability distribution, sample size, and test length. Ree (1979) also found that accuracy of estimation is dependent on the distribution of ability for certain sample sizes and

test lengths. Ree investigated the JML procedure with normal and truncated normal ability distributions. Ree's experiment was extended in this study by including a beta ability distribution and by including the MB and ML procedures. Differences were found not only among estimation procedures but also in the estimation accuracy obtained with the three ability distributions.

When both sample size and test length increased, estimates became more accurate, and, with some exceptions, negligible differences were observed among estimation procedures and among ability distributions. These exceptions are (a) for the JML estimates of a, b, and θ parameters with 60 items and 1000 examinees, appreciable differences were between the beta ability distribution and either the normal or the truncated normal ability distribution, (b) the MB and the MML estimates of the a and b parameters were more accurate with the normal than they were with the truncated normal or the beta ability distribution, (c) the MB and the MML estimates of the c parameters were more accurate with normal and beta ability distributions, and (d) the ML-MB and the ML-MML estimates were more accurate with the normal and the beta ability distribution than they were with the truncated normal distribution.

These results appear to be in disagreement with results of the study by Yen (1987) which indicated that the ability distribution did not affect estimation accuracy of a, b, c, and θ parameters. The apparent disagreement can be attributed to the differences in the conditions investigated in the two studies. The results from either of the two studies can only be generalized to situations similar to those investigated in the study. For example, if the a and the c parameters were constant and ability distribution slightly deviated from normality then it is more probable that estimation accuracy will not be affected by ability distribution. On the other hand, if a and c parameters were varied and the deviation from normality was more extreme than in the study by Yen, estimation accuracy will be affected by ability distribution as indicated by Ree (1979) and confirmed in the current study.

The contribution of this dissertation was to detect differences in accuracy among a broader array of estimation procedures than Ree (1979) investigated and in more realistic conditions than Yen (1987) investigated. Some differences prevailed even with long tests and large sample sizes, conditions that are favorable to accurate estimation with all three procedures.

The implications of finding differences in accuracy among estimation procedures and among ability distributions are

practical as well as theoretical. More accurate estimation occurred with certain distributions, estimation procedures, sample sizes, and test lengths. Recommendations for using BILOG or LOGIST can be based on these results. For example, the MB procedure is recommended when the sample size and/or test length is small because it results in estimates as accurate or more accurate than those produced by the other procedures. Nevertheless for some parameters the other procedures work nearly as well. The MB and the JML estimation procedures have similar accuracy for estimating the a parameters, MB and MML procedures have similar accuracy for estimating the b parameters except when θ is normally distributed, and the MB and JML procedures have similar accuracy for estimating the c parameters except with beta ability distribution. When guessing constitutes a problem of main concern and the ability distribution is beta the JML estimation is generally preferred.

APPENDIX A

FACTORS AND LEVELS IN THE CURRENT STUDY (1990)

Factors	Levels
Number of Items	20,60
Number of Examinees	250,1000
Parameter Distributions[a]	
θ	Standard Normal Truncated Normal Beta
a	Lognormal
b	Normal
c	Beta
Procedures	JML of LOGIST, and MB, MML, and ML of BILOG.

[a]The three parameter logistic model was used.

APPENDIX B

DERIVATION OF THE MEAN AND THE STANDARD DEVIATION OF THE TRUNCATED DISTRIBUTION

1. Derivation of the Mean

$$\mu = \frac{3}{2} \int_{-\infty}^{\infty} \frac{x}{2\pi^{\frac{1}{2}}} \quad e^{-\frac{1}{2}x^2} = \frac{3}{2(2\pi)^{\frac{1}{2}}} \left[\quad e^{-\frac{1}{2}c^2} \quad \right]$$

where c is the cutoff score of 0.053.

2. Derivation of the Standard Deviation

$$E(X^2) = \frac{3}{2} \int_{-\infty}^{c} \frac{x}{2\pi^{\frac{1}{2}}} \quad e^{-\frac{1}{2}x^2} \quad dx$$

$$= \frac{3}{2} \left[\int_{0}^{c} \frac{x}{2\pi^{\frac{1}{2}}} dx + \int_{-\infty}^{0} \frac{x}{2\pi^{\frac{1}{2}}} dx \right]$$

$$= \frac{3}{2} \left[\frac{1}{2\pi^2} \int_{0}^{c} \frac{x}{2\pi^{\frac{1}{2}}} dx + 0.5 \right]$$

$$= \frac{3}{2} \left[\frac{1}{2\pi^2} \int_{0}^{\frac{1}{2}c^2} \sqrt{2t} \quad e^{-t} \quad dt + 0.5 \right]$$

$$= \frac{3}{2}\left[\frac{T(3/2)}{\pi^{\frac{1}{2}}}\int_0^{\frac{1}{2}c^2}\frac{t^{3/2-1}\,e^{-t}}{T(3/2)}\,dt + 0.5\right]$$

$$= \frac{3}{2}\left[\frac{\pi^{\frac{1}{2}}}{2\pi^{\frac{1}{2}}}\int_0^{\frac{1}{2}c^2}\frac{t^{3/2}\,e^{-t}}{T(3/2)}\,dt + 0.5\right]$$

$$= \frac{3}{4}\left[\ IG\left(\tfrac{1}{2}c^2\ ;\ \alpha=\frac{3}{2}\ ,\ \beta=1\right) + 1\ \right]$$

where $t = \frac{1}{2}x^2$, $dt = xdx$, $x = \sqrt{2t}$, and $\sigma = \sqrt{\sigma^2}$.

BIBLIOGRAPHY

Birnbaum, A. (1957). Efficient design and use of tests of a mental ability of various decision making problems (Series Report No. 58-16. Project No. 7755-23). Randolph Air Force Base, TX: USAF School of Aviation Medicine, (School of Aviation Medicine).

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores (pp. 397-424). Reading, MA: Addison-Wesley.

Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R.D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored test items. Psychometrika, 35, 179-197.

Dempster, A.P., Rubin, D.B., & Tsutakawa R. K. (1981). Estimation in covariance components models. Journal of the American Statistical Association, 76, 341-353.

Divgi, D.R. (1985). A minimum Chi-square method for developing a common metric in item response theory. Applied Psychological Measurement, 9, 413-415.

Haberman, S. (1975). Maximum likelihood estimates in exponential response models. The Annals of Statistics, 5, 814-841.

Hambleton, R.K. & Cook, L.L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. Wiess (Ed.), New horizons in testing (pp. 33-48). New York, NY: Academic Press.

Hambleton, R.K., Murray, L.N., & Williams, P. (1983). Fitting item response models to the Maryland functional reading tests (Laboratory of Psychometric and Evaluative Research Report No. 139). Amherst, MA: School of Education, University of Massachusetts. (ERIC Document Reproduction Service No. ED 230 624)

Hambleton, R.K. & Rovinelli, R. (1973). A FORTRAN IV program for generating examinee response data from logistic test models. Behavioral Science, 18, 74.

Hambleton, R.K. & Swaminathan, H.H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff Publishing.

Hattie, J.A. (1984). An empirical study of various indices of determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982). Recovery of two- and three-parameter item characteristic curves: A Monte Carlo study. Applied Psychological Measurement, 6, 249-260.

Jensema, C.J. (1976). A simple technique for estimating latent trait mental test parameters. Educational and Psychological Measurement, 36, 705-715.

Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.L. (1982). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Lord, F.M. (1974). Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 39, 247-264.

Lord, F.M. (1975). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters (Research Bulletin 75-33). Princeton, NJ: Educational Testing Service.

Lord, F.M. (1977). Practical application of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F.M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. Journal of Educational Measurement, 23, 157-162

Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Marco, G.L. (1977). ICC solutions to three interactable testing problems. Journal of Educational Measurement, 14, 139-160

Mislevy, R.J. (1984). Estimating latent distributions. Psychometrika, 49, 359-381.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. Psychometrika, 51, 177-195.

Mislevy, R.J. & Bock, R.D. (1984). BILOG Version 2.2.: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

Mislevy, R.J. & Stocking, M. L. (1987). A consumer's guide to LOGIST and BILOG. Princeton, NJ: Educational Testing Service.

Pearson, E.S. & Hartly, H.O. (1956). Biometrika tables for statisticians (2nd ed). London: Cambridge University Press.

Qualls, A.L. & Ansley, T.N. (1985). A comparison of item and ability parameter estimates derived from LOGIST and BILOG. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Ree, M.J. (1979). Estimating item characteristic curves. Applied Psychological Measurement, 3, 371-385.

Rigdon, S.E. & Tsutakawa, R.K. (1981). Estimation in latent trait models (Research Report 81-1). Columbia, MO: University of Missouri. (ERIC Document Reproduction Service No. ED 208 033)

Rigdon, S.E. & Tsutakawa, R.K. (1983). Parameter estimation in latent trait models. Psychometrika, 48, 567-574.

Rigdon, S.E. & Tsutakawa, R.K. (1987). Estimation of the Rasch model when both ability and difficulty parameters are random. Journal of Educational Statistics, 12, 76-86.

Samejima, F. (1986). Results of item parameter estimation using LOGIST 5 on simulated data Knoxville, TN: Tennessee University. (ERIC Document Reproduction Service No. ED 265 213)

Stocking, M.L. & Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Swaminathan, H. & Gifford, J.A. (1982). Bayesian estimation in the Rasch model. Journal of Educational Statistics, 7, 175-191.

Swaminathan, H. & Gifford, J.A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. Weiss (Ed.), New horizons in testing (pp. 14-30). New York, NY: Academic Press.

Swaminathan, H. & Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model. Psychometrika, 50, 349-364.

Swaminathan, H. & Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. Psychometrika, 51, 589-601.

Swaminathan, H. & Gifford, J.A. (1987). A comparison of the joint and marginal maximum likelihood procedures for the estimation of parameters in item response models. Draft final report submitted to the Institute for Student Assessment and Evaluation, University of Florida, Gainesville, FL.

Tsutakawa, R.K. (1984). Improved estimation procedures for item response functions (Final Report on Project NR150-464. Research Report 84-2). Columbia, MO: University of Missouri. (ERIC Document Reproduction Service No. ED 250 397)

Tsutakawa, R.K. & Lin, H.Y. (1986). Bayesian estimation of item response curves. Psychometrika, 51, 251-267.

Urry, V.W. (1977). Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 14, 181-196.

Vale, C.D. & Gialluca, K.A. (1985). ASCAL: A microcomputer program for estimating logistic IRT item parameters (Research Report ONR-85-4). St. Paul, MN: Assessment Systems Corporation.

Vale, C.D. & Gialluca, K.A. (1988). Evaluation of the efficiency of item calibration. Applied Psychological Measurement. 12, 53-67.

Wingersky, M.S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R.K. Hambleton (Ed.), Applications of item response theory (pp. 151-156). Vancouver, B.C.: Educational Research Institute of British Columbia.

Wingersky, M.S., Barton, M.A., & Lord F.M. (1982; Version 2.5 updated 1984). LOGIST 5.0 version 1.0 users' guide. Princeton, NJ: Educational Testing Services.

Wingersky, M.S. & Lord, F.M. (1973). A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses (RM-73-2). Princeton, NJ: Educational Testing Services.

Wingersky, M.S. & Lord, F.M. (1984). An Investigation of Methods for Reducing Sampling Error in Certain Item Response Theory Procedures. Applied Psychological Measurement, 8, 347-364.

Yen, W.M. (1983). Tau equivalence and equipercentile equating. Psychometrika, 48, 353-369.

Yen, W.M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. Journal of Educational Measurement, 21, 93-111.

Yen, W.M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.

BIOGRAPHICAL SKETCH

Abdel-fattah A. Abdel-fattah was born in 1955. He received a Bachelor of Arts in zoology from the University of Ain-Shams of Egypt, in 1978. In 1982 he began work towards the Ph.D degree at the University of Florida.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

James Algina

James Algina, Chairman
Professor of Foundations of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Linda Crocker

Linda M. Crocker
Professor of Foundations of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Fatimah Linda Collier Jackson

Fatimah Linda Collier Jackson
Associate Professor of Anthropology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

David Miller

David Miller
Assistant Professor of Foundations of Education

This dissertation was submitted to the Graduate Faculty of the College of Education and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 1990

James Algina
Chair, Foundations of Education

David C. Smith
Dean, College of Education

Dean, Graduate School